DOCUMENT RESUME

ED 406 417 TM 026 339

AUTHOR	Martin, Michael O., Ed.; Mullis, Ina V. S., Ed.
TITLE	Third International Mathematics and Science Study: Quality Assurance in Data Collection.
INSTITUTION	Boston Coll., Chestnut Hill, MA. Center for the Study of Testing, Evaluation, and Educational Policy.; International Association for the Evaluation of Educational Achievement.
REPORT NO	ISBN-1-889938-01-7
PUB DATE	96
NOTE	286p.; Funding for the international coordination of the Third International Mathematics Study is provided by the National Center for Education Statistics, the National Science Foundation, the International Association for the Evaluation of Educational Achievement, and the Canadian government.
AVAILABLE FROM	Boston College, Center for the Study of Testing, Evaluation, and Educational Policy, Campion Hall 323, Chestnut Hill, MA 02167; http://wwwcsteep.bc.edu/timss
PUB TYPE	Collected Works - General (020) -- Reports - Evaluative/Feasibility (142) -- Statistical Data (110)
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	*Academic Achievement; Comparative Analysis; *Data Collection; Elementary Secondary Education; Foreign Countries; International Education; *International Studies; *Mathematics Tests; *Quality Control; Questionnaires; Reliability; Research Methodology; Sampling; *Science Tests; Scoring; Tables (Data); Testing; Translation
IDENTIFIERS	International Assn Evaluation Educ Achievement; Monitoring; *Third International Mathematics and Science Study

ABSTRACT

The Third International Mathematics and Science Study (TIMSS) is the most ambitious study conducted by the International Association for the Evaluation of Educational Achievement to date. TIMSS developed and administered tests and questionnaires in three student populations to study achievement in participating countries and the factors associated with high achievement. Contributions to this volume describe the main elements of the effort to assure the quality of TIMSS data, focusing on such issues as the translation of the achievement tests from English into the 30 other languages of the study, sampling from the 45 participating countries, and reliability questions. The following chapters are included: (1) "Translation, Verification" (Ina V. S. Mullis, Dana L. Kelly, and Kathleen Haley); (2) "Sampling" (Pierre Foy, Michael O. Martin, and Dana L. Kelly); (3) "Monitoring the TIMSS Data Collection" (Michael O. Martin, Craig D. Hoyle, and Kelvin D. Gregory); (4) "Observing the TIMSS Test Administration" (Michael O. Martin, Craig D. Hoyle, and Kelvin D. Gregory); (5) "Quality Control Steps for Free-Response Scoring" (Ina V. S. Mullis and Teresa A. Smith); and (6) "Data Consistency Checking across Countries" (Heiko Jungclaus and Michael Bruneforth). Nine appendixes present supplemental information about study methodology. (Contains 1 figure, 28 tables, and 3 appendix tables.) (SLD)

ED 406 417

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
MICHAEL O. MARTIN
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Third International Mathematics and Science Study: Quality Assurance in Data Collection

TM026339

IEA
TIMSS

BEST COPY AVAILABLE

TIMSS

Quality Assurance in Data Collection

Third International Mathematics and Science Study

Quality Assurance in Data Collection

Edited by
Michael O. Martin
Ina V.S. Mullils

with contributors
Michael Bruneforth
Pierre Foy
Kelvin D. Gregory
Kathleen Haley
Craig D. Hoyle
Heiko Jungclaus
Dana L. Kelly
Teresa A. Smith

Boston College • Chestnut Hill, Massachusetts

© 1996 International Association for the Evaluation of Educational Achievement (IEA).

Third international mathematics and science study: quality assurance in data collection/edited by Michael O. Martin, Ina V.S. Mullis
Publisher: Center for the Study of Testing, Evaluation, and Educational Policy,
Boston College.

Library of Congress Catalog Card Number: 96-71249

ISBN 1-889938-01-7

For more information about TIMSS contact:

TIMSS International Study Center
Center for the Study of Testing, Evaluation, and Educational Policy
Campion Hall
School of Education
Boston College
Chestnut Hill, MA 02167
United States

This report also is available on the World Wide Web:
http://wwwcsteep.bc.edu/timss

Boston College is an equal opportunity/affirmative action employer.
Funding for the international coordination of TIMSS is provided by the U.S. National Center for Education Statistics, the U.S. National Science Foundation, the IEA, and the Canadian government. Each participating country provides funding for the national implementation of TIMSS.

Printed and bound in the United States.

CONTENTS

FOREWORD vii

ACKNOWLEDGMENTS ix

INTRODUCTION 1

1. TRANSLATION VERIFICATION 1-1

Ina V.S. Mullis, Dana L. Kelly, and Kathleen Haley

1.1 OVERVIEW 1-1
1.2 TRANSLATION GUIDELINES 1-3
1.3 ADAPTATION REQUESTS 1-4
1.4 TRANSLATION VERIFICATION 1-5
1.5 VERIFICATION BY THE TIMSS QUALITY CONTROL MONITORS 1-8
1.6 VERIFICATION BY THE INTERNATIONAL STUDY CENTER 1-8
1.7 STATISTICAL AND CONTENT-RELATED CONSIDERATIONS 1-10
1.8 SUMMARY 1-11

2. SAMPLING 2-1

Pierre Foy, Michael O. Martin, and Dana L. Kelly

2.1 OVERVIEW 2-1
2.2 DOCUMENTATION OF THE SAMPLING PROCEDURES 2-2
2.3 POPULATION DEFINITIONS AND SAMPLE PARTICIPATION RATES 2-11
2.4 REPORTING ACHIEVEMENT RESULTS 2-21
2.5 SUMMARY 2-22

3. MONITORING THE TIMSS DATA COLLECTION 3-1

Michael O. Martin, Craig D. Hoyle, and Kelvin D. Gregory

3.1 THE TIMSS QUALITY CONTROL MONITORS 3-1
3.2 TRAINING OF QUALITY CONTROL MONITORS 3-1
3.3 THE QUALITY CONTROL MONITOR'S VISIT TO THE NATIONAL CENTER 3-3
3.4 SUMMARY OF RESULTS OF INTERVIEWS WITH NRCS 3-3
3.5 SELECTION OF SCHOOLS FOR CLASSROOM OBSERVATION 3-8
3.6 NATIONAL VERSIONS OF DATA COLLECTION INSTRUMENTS 3-9
3.7 SURVEY ACTIVITIES REPORT 3-9
3.8 SUMMARY 3-12

4. OBSERVING THE TIMSS TEST ADMINISTRATION 4-1
Michael O. Martin, Craig D. Hoyle, and Kelvin D. Gregory

4.1 OVERVIEW 4-1
4.2 SCHOOL VISITS AND TEST SESSION OBSERVATIONS 4-3
4.3 SUMMARY 4-10

5. QUALITY CONTROL STEPS FOR FREE-RESPONSE SCORING 5-1
Ina V.S. Mullis and Teresa A. Smith

5.1 OVERVIEW 5-1
5.2 TRAINING SESSIONS FOR FREE-RESPONSE SCORING 5-2
5.3 WITHIN-COUNTRY RELIABILITY STUDIES 5-4
5.4 IMPLEMENTING THE CROSS-COUNTRY RELIABILITY STUDY 5-9
5.5 THE RESULTS OF THE INTERNATIONAL CROSS-COUNTRY CODING STUDY 5-14
5.6 SUMMARY 5-29

6. DATA CONSISTENCY CHECKING ACROSS COUNTRIES.......6-1
Heiko Jungclaus and Michael Bruneforth

6.1 OVERVIEW 6-1
6.2 DATA CLEANING 6-1
6.3 IDENTIFICATION VARIABLES, TRACKING VARIABLES, AND INDICATORS 6-5
6.4 PRELIMINARY STATISTICS 6-7
6.5 DATABASE CONSTRUCTION 6-8
6.6 SUMMARY 6-9

APPENDIX A **Examples of Item Information**

APPENDIX B **Changes Made to the TIMSS International Database for the Population 2 Cognitive Items**

APPENDIX C **TIMSS Sampling Forms**

APPENDIX D **TIMSS Quality Control Monitors**

APPENDIX E **Results of the Quality Assurance Monitors' Interviews with the National Research Coordinators**

APPENDIX F **Results of the Quality Assurance Monitors' Test Session Observations**

APPENDIX G **TIMSS International Coding Reliability Study Participants**

APPENDIX H **Contingency Tables and Coding Guides for Items in the International Reliability Study**

APPENDIX I **Data Cleaning and Consistency Checks**

Foreword

The Third International Mathematics and Science Study (TIMSS) is the most ambitious study conducted by the International Association for the Evaluation of Educational Achievement (IEA) to date. TIMSS developed and administered tests and questionnaires in three student populations with the aim of studying student achievement in mathematics and science in participating countries, and the home, school, and instructional factors associated with high performance. The design of the study, the development of the instruments and procedures, and the data collection, were achieved through a collaborative process that involved hundreds of individuals around the world.

In a study as large and complex as TIMSS, it was critical that standardized data collection procedures be developed and implemented to ensure that comparable high-quality data were collected in each country. With the encouragement and generous financial support of the U. S. National Center for Education Statistics, and the National Science Foundation, the staff of the TIMSS International Study Center developed a program of quality assurance initiatives in furtherance of that aim. In this volume, Michael Martin and Ina Mullis have described the main elements of that program, specifying the procedures undertaken to ensure high-quality TIMSS data and the success with which those procedures were implemented. The report pays particular attention to issues such as the translation of the achievement tests from English to the other 30 languages of the study, the sampling of student populations in the 45 participating countries, the administration of the tests and questionnaires in participating classrooms and schools, the reliability of the scoring of the large numbers of free-response achievement items in the tests, and the consistency and integrity of the database.

TIMSS plans to publish its first international reports containing achievement results (for seventh and eighth grades) in November 1996, with further reports (for third and fourth grades, and for twelfth grade) to follow in 1997. This volume will help readers of the international reports to appreciate the efforts to achieve high-quality data in TIMSS and to evaluate the success of those endeavors.

ALBERT E. BEATON
International Study Director
Boston College

Acknowledgments

TIMSS was truly a collaborative effort among hundreds of individuals around the world. Staff from the national research centers of the participating countries, the international management, advisors, and funding agencies worked closely to design and implement the most ambitious study of international comparative achievement in mathematics and science ever undertaken. The TIMSS quality assurance efforts were also accomplished through a collaborative effort. National Research Coordinators (NRCs) worked to ensure a high-quality data collection by carefully and conscientiously following the internationally prescribed procedures, sometimes in the face of considerable difficulties and with minimal resources. NRCs also made a vital contribution to the success of the program of visits by quality assurance monitors by nominating suitable monitors, participating in detailed interviews, and helping to organize school visits.

Eugene Owen and Lois Peak from the U. S. National Center for Education Statistics and Larry Suter from the National Science Foundation provided encouragement and support at all stages of the endeavor.

Jules and Marlene Goodison helped to design the training manuals for quality control monitors and to organize and carry out the training program. They also helped to design the data collection instruments for the visits to NRCs and for classroom observations. Ken Ross and Hans Pelgrum reviewed the materials and offered many useful suggestions. The quality control monitors (listed in Appendix D) played a crucial role in monitoring adherence to standardized procedures in each country.

This report would not have been possible without the efforts of many people. TIMSS International Study Director Albert Beaton was a never-ending source of support and wisdom. As Director of Operations and Data Analysis, Eugenio Gonzalez played a major role in designing and supporting the quality assurance program, and provided data analysis services at important stages of the process. Russell Jones contributed significantly to the management and analysis of the data from the reports of the quality assurance monitors. Each of the chapter authors contributed not only to their own chapter, but also provided thoughtful reviews of the entire volume.

Several people at the TIMSS International Study Center, Boston College, deserve special acknowledgment for the production of this report. Dana Kelly supervised the design and development of the report. Maria Sachs edited the early drafts, and provided many helpful suggestions. Cheryl Flaherty managed the details of putting the report together. She tirelessly and cheerfully edited and revised numerous drafts of the report, and ensured that all pieces were in the right place. Kelvin Gregory and Craig Hoyle read and commented on the successive drafts. Jon English created many of the data displays and assisted in the layout. Debora Galanti designed the cover and Tom Hoffmann coordinated production of the document.

MICHAEL O. MARTIN
INA V.S. MULLIS
Boston College

INTRODUCTION

The Third International Mathematics and Science Study (TIMSS), a comparative study of student achievement in mathematics and science, is a huge, complex project involving 45 countries, three student populations incorporating five grade levels, and over half a million students. Although the study is directed from the International Study Center[1] at Boston College, each participating country was responsible for implementing the design in that country in accordance with the international standards. Survey instruments and field procedures were developed through a process of cooperation and consensus among the participants, and fieldwork was carried out by the National Research Coordinator (NRC) in each country. Each participating country or educational system was responsible for translating the instruments and procedures and adapting them to local conditions, drawing the school and student samples, and implementing the data collection plan. This combination of international cooperation and national implementation is an efficient and cost-effective approach to conducting international comparative studies, but it requires close collaboration among participants, and the validity and reliability of the results are crucially dependent on each participant adhering to the prescribed procedures at all times.

TIMSS has expended considerable effort in developing standardized materials and procedures so that the data collected in all countries are comparable to the greatest possible extent. Martin, Mullis, and Kelly (1996) have documented these efforts, which include the provision of extensive technical documentation, translation verification procedures, training seminars, individual consultation, and computer software. It is important not only that the TIMSS data be of high quality but that the project be able to demonstrate the quality of the data to readers and users of the TIMSS reports and data. Conscious of this and of criticisms that have been made of some procedures used in previous international studies, the main funders of the International Study Center, the U.S. National Center for Educational Statistics and the U.S. National Science Foundation, provided additional funds for a quality assurance program to document the quality of the TIMSS data in several areas that have been subject to criticism in past studies.

The quality assurance program specified a range of activities spread over a three-year period. The aim was to help ensure the comparability of results across participating countries, and to provide documentation to assist in the interpretation of the data. The main activities are described in this report. They were designed to enhance and document the quality of the TIMSS data, with particular emphasis on instrument translation and adaptation, sampling response rates, test administration and data collection, the reliability of the coding process, and the integrity of the database. The chapters dealing with the preparations for data collection, the data collection itself, and the checking and processing of the data (Chapters 1, 3, 4, and 6) refer to all three student populations. However, those chapters that report results from the data (Chapter 2 on sampling and Chapter 5 on the reliability of the coding process) are restricted to Population 2,[2] which was compulsory for all participants.

---

[1] The study was coordinated from its inception until August 1993 by the International Coordinating Center (ICC) at the University of British Columbia, Vancouver, Canada. From August 1993, the study was directed from the International Study Center at Boston College.

In any comparative study of student achievement that takes place in more than one language there is a risk that the difficulty of the tests (which are usually constructed in one language) may be affected by translation into other languages. With the administration of tests and questionnaires in 31 languages, this was an issue of great concern for TIMSS. Maxwell (1996) describes the procedures that were developed to assist NRCs in producing high-quality translations, and to monitor the quality of the translation process. In Chapter 1 of this report, Mullis, Kelly, and Haley review the translation verification procedures, and report on the status of the translation effort.

International comparative studies like TIMSS, which seek to make inferences about national populations on the basis of sample survey methodology, rely on the quality of the national samples for the validity of those inferences. In TIMSS great attention was paid to all aspects of the population sampling process, from population definition through sample design and selection to computation of participation rates, sampling weights, and estimates of sampling variance. Foy, Rust, and Schleicher (1996) describe the sampling design in detail. In Chapter 2 of the present report, Foy, Martin, and Kelly document participants' compliance with prescribed procedures at each stage of the sampling process, and present data on population coverage and participation for each participant.

The TIMSS achievement tests were designed to be administered under uniform conditions throughout all participating countries. Documenting the uniformity of the test administration required that a sample of testing sessions in each country be observed. In order to visit schools and carry out such observations it was necessary to hire and train a quality control monitor for each country. Monitors had to be fluent both in English (the language of the training and monitoring materials) and in the language of the country. They had two major tasks. The first was to visit the TIMSS national center to interview the NRC about all aspects of the data collection, including sampling, instrument translation, production and shipping, and plans for receipt control, free-response coding, and data entry. The International Study Center prepared data collection instruments and a manual to be used by the quality control monitors, and organized regional training meetings to ensure that the monitors were well versed in all of their responsibilities. Martin, Hoyle, and Gregory, in Chapter 3, describe this activity in detail, including the development of data collection instruments for the visits and the design and implementation of a training program for the quality control monitors. This chapter also includes a summary of the results of the interview with NRCs.

---

[2] Population 2 is defined as the two adjacent grades with the largest proportion of 13-year-old students at the time of testing.

The second task of the quality control monitor was to visit a random selection of schools from the TIMSS sample at the time testing was taking place and determine whether the tests were being administered using uniform and secure procedures. Each quality control monitor was required to visit ten schools in the TIMSS sample, to observe a testing session, and to interview the school coordinator regarding the implementation of the TIMSS procedures. These school visits and test session observations are a central component of the TIMSS quality assurance effort. The quality control monitor completed one classroom observation record for each visit. Martin, Hoyle, and Gregory summarize the results of these observations in Chapter 4, and present the results in detail in an accompanying appendix.

The TIMSS achievement tests included both multiple-choice and free-response (open-ended) items. Many of the free-response items required an extended response from students, and all of them required that the student responses be coded by trained coders prior to data entry. Detailed coding rubrics with example codes were provided for each item, and regional training meetings were organized by the International Study Center to ensure that each participant had a full understanding of the application of the rubrics. To monitor the reliability of the coding process in each country, each participant was required to select a 10% random sample of student responses and code them twice, using different coders on each occasion. Reliability coefficients were computed for each item in every country that complied with this requirement. In order to provide an indication of the consistency of coding across countries, English-speaking coders from 21 countries came to a central location and coded samples of student responses from seven English-speaking countries. In Chapter 5, Mullis and Smith present the results of the reliability studies both within and across countries, and discuss their significance for the quality of the TIMSS data.

Accurate and reliable comparisons of international achievement require accurate and complete datasets from participating countries. Although each participant was responsible for coding, entering, and checking that country's data, and for ensuring that all data were in the prescribed international format, the IEA Data Processing Center (DPC) in Hamburg, Germany, was charged with verifying that participants had complied with the international standard. The scale and complexity of the TIMSS tests and questionnaires required an enormous data verification exercise. The quality assurance program provided support for the staff of the DPC as they engaged in an extensive series of quality control checks and communicated with each country regarding the nature and extent of the national deviations from prescribed international procedures. In Chapter 6, Jungclaus and Bruneforth describe the procedures used to verify the data and the actions taken to remedy any deviations.

The activities described in this report should provide assurance to readers of TIMSS publications and users of TIMSS data that the highest professional standards were applied in all phases of the data collection, and that a very high standard was attained in all stages of the endeavor.

REFERENCES

Martin, M.O., Mullis, I.V.S., and Kelly, D.L. (1996). "Quality Assurance Procedures" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Maxwell, B. (1996). "Translation and Cultural Adaptation of the Survey Instruments" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Foy, P., Rust, K., and Schleicher, A. (1996). "Sample Design" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

1. TRANSLATION VERIFICATION PROCEDURES

Ina V.S. Mullis
Dana L. Kelly
Kathleen Haley

1.1 OVERVIEW

The TIMSS instruments were prepared in English and translated into 30 additional languages across the 45 participating countries (see Table 1.1, below, for the list of languages). In addition, it sometimes was necessary to adapt the international, that is, original, versions for cultural purposes, even for the 11 countries that tested in English. To ensure the standardization of the instruments across languages and countries, and thus the comparability of the TIMSS data, explicit guidelines for translation and cultural adaptation were developed and comprehensive verification procedures were implemented. Specifically, the TIMSS instrument translation effort included the development of explicit guidelines for translation and cultural adaptation; translation of the instruments by the national centers in accordance with the guidelines; verification of the quality of the translations and booklet layout by independent translators; verification by quality control monitors that suggested changes to the translations (if any) were made by the national centers; and a series of statistical checks after the testing to detect items that did not perform comparably across countries.

International versions of the data collection instruments were produced centrally and translated by the national centers in accordance with the translation guidelines.[1] The International Study Center provided each national center with paper and electronic versions of the cognitive items, together with fully assembled booklets in paper format. The questionnaires were provided both in electronic and paper versions. Explicit instructions were provided for translation and cultural adaptation of the cognitive items and assembly and layout of the test booklets and questionnaires. In addition, the questionnaires were accompanied by instructions for adapting certain questionnaire items. Each national center was responsible for producing the translated instruments using the translation guidelines provided and for submitting the translated, camera-ready versions of the instruments to the International Study Center for verification.

---

[1] The International Coordinating Center (ICC) produced the item booklets for the 1993 item pilot and for the 1994 field trial; the International Study Center produced the instruments for the main survey.

Table 1.1
Languages in Which the TIMSS Instruments Were Administered

Language	Countries
Afrikaans	South Africa
Arabic	Kuwait
Bulgarian	Bulgaria
Chinese	Hong Kong
Czech	Czech Republic
Danish	Denmark
Dutch	The Netherlands
English	Australia, Canada, England, Hong Kong, Ireland, New Zealand, Philippines, Scotland, Singapore, South Africa, United States
Farsi	Iran
Flemish	Belgium
French	Belgium, Canada, France, Switzerland
German	Austria, Germany, Switzerland
Greek	Cyprus, Greece
Hebrew	Israel
Hungarian	Hungary
Icelandic	Iceland
Indonesian	Indonesia
Italian	Italy, Switzerland
Japanese	Japan
Korean	Korea
Latvian	Latvia
Lithuanian	Lithuania
Norwegian	Norway
Portuguese	Portugal
Romanian	Romania
Russian Federation	Russian Federation
Slovak	Slovak Republic
Slovene	Slovenia
Spanish	Argentina, Colombia, Mexico, Spain
Swedish	Sweden
Thai	Thailand

1.2 TRANSLATION GUIDELINES

In 1992, the International Coordinating Center (ICC) drafted guidelines for translation and cultural adaptation of the TIMSS tests based on a paper prepared for TIMSS by Ronald K. Hambleton (1992) regarding the translation of achievement tests into multiple languages. These guidelines were developed to ensure that the cognitive items would be translated from the international versions into the target languages without changes in meaning or difficulty; that cultural differences would be kept to a minimum; and that the meaning and content of the questionnaire items would be retained through translation. The goal was to obtain translated instruments of high quality that would provide comparable data across countries and cultures.

The translation guidelines developed for TIMSS and described below are documented in the *Survey Operations Manuals* (TIMSS, 1994a, 1994b). They recommended that each national center engage a minimum of four translators–two mathematics and two science education specialists with fluency in English and the target language. Only two translators were necessary for the translation of the questionnaire items. The mathematics specialists were to produce two independent translations of the mathematics cognitive items, and the two science specialists were to produce two independent translations of the science cognitive items.

In general, the translators' work included the following:

- Identifying and minimizing cultural differences
- Finding equivalent words and phrases
- Making sure the reading level was the same in the target language as in the international version
- Making sure the essential meaning of the items did not change
- Making sure the difficulty of the achievement items did not change
- Being aware of changes in layout due to translation.

The *Survey Operations Manuals* (TIMSS, 1994a, 1994b) also provide guidelines regarding decisions about vocabulary, meaning, layout, and cultural adaptations. These guidelines include examples of acceptable and unacceptable cultural adaptations to the items, such as changes in punctuation, units, proper nouns, common nouns, spelling, verbs (not related to content), and usage. Furthermore, translators were instructed that when modifying the text of an item for cultural adaptation purposes the following were to remain the same as the international version:

- The meaning of the question
- The reading level of the text
- The difficulty of the item
- The likelihood of another possible correct answer for the test item.

Each pair of translators (for mathematics and for science) was to translate the test items independently and then have the two versions compared by a third party. When there were differences between the two versions, the best version of the translation was selected. Any deviations in vocabulary, meaning, or item layout from the international versions were recorded on the Translation Deviation Form. The completed forms were then submitted to the International Study Center together with the translated instruments, and used during the international translation review.

Countries were also encouraged to send their suggestions for adaptations to the International Study Center as they were translating the items. This procedure, described further in section 1.3, allowed the national centers to receive approval of adaptations quickly, which expedited the preparation of the test booklets.

Due to limited resources, some countries were unable to engage more than one translator each for mathematics and science. However, by engaging translators with the appropriate and recommended qualifications and by adhering to the international procedures for translation, countries were still able to produce high-quality translated instruments which were commensurate with the international versions.

1.3 ADAPTATION REQUESTS

While the intention of TIMSS was to devise items that would be interpreted similarly across countries, it was clear that a single version, strictly translated, would not be wholly appropriate in every country. Therefore, a procedure was put in place for requesting approval for specific adaptations to the items. This process was distinct from the translation verification process in that it allowed an NRC to obtain timely approval for particular adaptations, usually within two days. This expedited the production of the test booklets at the national centers and brought any deviations to the attention of the subject area coordinators and the International Study Center prior to the printing and duplication of the booklets at the national centers.

In order to gain approval for an adaptation, the NRC documented the request on the Item Adaptation Request Form and sent it to the International Study Center. This took place either during the national translation and booklet production effort or after submission of testing materials for translation verification. Upon receipt of the request forms at the International Study Center, the adaptations were forwarded to the appropriate subject area coordinator for review. The math or science coordinator considered each request and approved the change, rejected it, or specified additional modifications necessary to make it acceptable. Staff at the International Study Center were responsible for informing the national centers whether the requested adaptations were approved.

Item adaptations were approved if they did not in any way change the substance or intent of the question or answer choices. For example, a change from "weight" to "weight (mass)" was an acceptable clarification for students unaccustomed to the colloquial use of "weight," while simply changing "weight" to "mass" would be unacceptable as it would make the units inappropriate. Similarly, requests from a number of countries to replace "congruent" with "same shape and size" were approved, whereas the use of "equal" instead of "congruent" was regarded as too imprecise and therefore was not approved.

1.4 TRANSLATION VERIFICATION

The items administered in the 1993 item pilot were translated by the national centers according to the guidelines described in section 1.2 and submitted to the ICC for verification. The translation verification was conducted after the items had been administered, the results had been analyzed, and items that were potential candidates for the field trial had been identified. Each NRC received a report on the quality of the translation of each of those items and had the opportunity to improve the translation, if necessary, for the field trial.

In the 1994 field trial, a slightly different approach was taken in order to conserve resources. If a country's item pilot translations had been deemed acceptable by the internationally commissioned translators, a 25% sample of the items was reviewed (following administration) to verify that the same quality existed in the field trial instruments. If a country's item pilot translations were not acceptable, or if that country did not participate in the pilot study, then all of the items were checked for the quality of the translation (Maxwell, 1996).

In the main survey, procedures required that the translated items and camera-ready copies of the test booklets be verified prior to printing and administration of the tests. Following translation and assembly of the test booklets, countries were required to send the items and the assembled booklets for each population that participated to the International Study Center. Additionally, the Population 2 performance assessment student booklets were verified. By having the translations verified prior to duplication, any errors or deviations found in the international review could be corrected. Due to the tight timeline between the completion of the international versions of the instruments and the main survey administration, some countries were not able to have their translations verified prior to printing. In these cases, efforts were made to ensure that the instruments were nonetheless verified.

Once the translated items, camera-ready test booklets, and completed Translation Deviation Forms were received by the International Study Center, they were forwarded to the ICC in Vancouver, Canada, for review by professional translators.[2] When NRCs included Item Adaptation Request Forms, the International Study Center forwarded the requests to the subject area coordinators for review. The translation of the items and booklet layout was checked by translators from the same professional translation company that had completed the verification of the instruments in the item pilot and field trial. The agency is based in Vancouver. Beverley Maxwell of the ICC coordinated the international review. She communicated with the translators regarding procedures, reviewed translators' reports and suggestions, and forwarded the translation verification reports to the national centers and the International Study Center. All those who verified the national translations had formal credentials as translators into the target language, first-language experience in the target language, excellent knowledge of English, experience living and working in an English-language environment, and familiarity with the culture associated with the target language.

---

[2] Translators reviewed instruments translated into languages other than English; booklets adapted by English-testing countries were reviewed by staff at the ICC.

Translators were provided with a number of materials to aid them in their understanding of the translation procedures used by the national centers and with instructions to carry out the review of the instruments. These materials are listed below (excerpted from Maxwell, 1996).

- A two-page introduction summarizing the TIMSS project, the instruments, and the translation goals, as background information
- A set of the translated instruments (as either booklets or clusters)
- A set of the international originals
- A copy of "Guidelines for Translation and Cultural Adaptation" (an excerpt from the *Survey Operations Manual* (TIMSS, 1994a, 1994b) containing the original instructions for translating the instruments; this allowed the verifier to know what instructions were given to the original translator)
- Instructions for verifying the general layout (checking that the message to students appeared at the beginning of the book, the questions appeared in the correct order, the illustrations were in the right place, all labels were translated, and page breaks were the same as in the international versions)
- Instructions for verifying the message to students (a list of points that the message must have clearly communicated)
- Instructions for item-by-item checking (including the procedures for coding observations to indicate the type and severity of the error)
- An example of a verified translation, including an annotated verifier's report.

For each country, a translator reviewed the overall layout of the instruments, the translation of the student instructions, and the translation of each item. The translator compared each translated item with the international version and documented any adaptations in a translation verification report. The translator assigned to each item that differed from the international version a code for the *type* of deviation and a code for the *severity* of deviation. The translator further provided an explanation of the change and how it could be corrected or improved upon, if necessary. The "type codes" and "severity codes" are described below.

TYPE CODES

The type codes, listed below, indicate what kind of change was made in the translation from the international version to the target language. Codes A through J refer to deviations in the text of an item; K through N refer to deviations in the graphics or layout of an item.

The type codes are:

A	Spelling
B	Grammar
C	Vocabulary
D	Incorrect number or value
E	Error in equation or numeric notation
F	Missing or additional text
G	Change in meaning
H	Change in level of reading difficulty
I	Tabs, alignment, or text layout
J	Other problem with the text
K	Labels are missing
L	Wrong picture or picture is missing
M	Picture has been modified
N	Labels have been modified.

SEVERITY CODES

The severity codes ranged from 1 (serious error) to 4 (acceptable adaptation).

1. Major Change or Error: This could affect the results and NRCs were to make corrections. Examples include incorrect ordering of choices in a multiple-choice item; omission of a graph that is essential to a solution; an incorrect translation of text such that the answer is indicated by the question.
2. Minor Change or Error: This was to be corrected if possible, but would not affect the results. Examples include spelling errors that do not affect comprehension; misalignment of margins or tabs; incorrect font or font size.
3. Suggestions for Alternative: The translation may have been adequate, but the verifier suggested a different wording for the item. The NRC was asked to review such suggestions and decide whether to make the suggested changes.
4. Acceptable Changes: The verifier identified acceptable changes and appropriate adaptations. This was done to provide information and required no action from the NRC. An example is where a reference to winter was changed from January to July for the Southern Hemisphere.

A code, comprised of the severity code and the type code, was assigned to each item for which the translator noted a deviation from the international version. For example, an appropriate change in vocabulary (coyote to dingo) would be coded as 4-C. An inappropriate change (gravity to weight) would be coded as 1-C. In cases where the verifier was unsure about the coding, a question mark was used in place of a code, and the uncertainty was elaborated upon in the explanation (Maxwell, 1996).

The translation verification report consisted of an overall statement of the quality of the translation, followed by a list of observations associated with individual items. The reports were sent to the ICC, where they were reviewed by Beverley Maxwell and subsequently forwarded to the International Study Center and the appropriate NRC. The above procedure required between four and six weeks; less time was required for instruments adapted for English-testing countries. Verification reports for all countries were entered into a database at the International Study Center.

1.5 VERIFICATION BY THE TIMSS QUALITY CONTROL MONITORS

When visiting the national center, the quality control monitor checked the final instruments against the translation verification report prepared by the independent reviewer to ensure that the suggestions for corrections and improvements had been followed. The quality control monitor recorded this in his/her report to the International Study Center.

1.6 VERIFICATION BY THE INTERNATIONAL STUDY CENTER

After translation verification, a further quality check of the instruments was made by the International Study Center. All deviations/adaptations coded 1 (major change or error) in the translation verification report were reviewed to determine whether a threat to validity existed. Final printed booklets were inspected to determine whether the error had been corrected. Between the verification conducted by the quality control monitors and that conducted by the International Study Center, it was determined that nearly all corrections to items coded as Type 1 had been made at Population 2. (Checking at Populations 1 and 3 is still in progress.) Table 1.2 presents for each country the number of translation deviations coded as Type 1 that were still present in the final Population 2 test booklets.

Table 1.2
Number of Potential Translation Errors, After Checking Final Test Booklets

Country	Population 1		Population 2		Population 3		
	Math	Science	Math	Science	Literacy	Advanced Math	Physics
Australia	0	0	0	0	0	0	0
Austria	1	0	0	0	0	0	0
Belgium (Fl)	0	0	0	0	–	–	–
Belgium (Fr)	–	–	*	*	–	–	–
Bulgaria	0	0	0	0	0	0	0
Canada (Eng)	0	0	0	0	0	0	0
Canada (Fr)	0	0	0	0	1	0	0
Colombia	–	–	0	0	–	–	–
Cyprus	0	0	0	2	1	1	0
Czech Republic	0	0	0	0	1	0	1
Denmark	–	–	0	0	0	0	0
England	0	0	0	0	–	–	–
France	0	–	0	0	0	0	0
Germany	–	–	*	*	*	*	*
Greece	0	0	0	0	*	*	*
Hong Kong	0	0	0	0	0	1	0
Hungary	0	0	1	0	*	–	–
Iceland	1	0	0	0	*	–	–
Indonesia	5	5	5	2	–	–	–
Iran	*	*	*	*	–	–	–
Ireland	0	0	0	0	–	–	–
Israel	0	0	0	1	*	*	*
Japan	0	1	0	0	–	–	–
Korea	*	*	*	*	–	–	–
Kuwait	1	1	5	1	–	–	–
Latvia	0	0	3	0	–	–	0
Lithuania	–	–	0	0	1	0	–
Mexico	0	0	0	0	1	*	*
Netherlands	0	0	0	0	0	–	–
New Zealand	0	0	0	0	0	0	0
Norway	0	0	0	0	0	–	0
Philippines	–	–	0	0	–	–	–
Portugal	0	0	0	0	–	–	–
Romania	–	–	1	0	–	–	–
Russia	–	–	*	*	*	*	*
Scotland	0	0	0	0	–	–	–
Singapore	0	0	0	0	–	–	–
Slovak Republic	–	–	0	1	–	–	–
Slovenia	0	0	0	0	0	2	0
South Africa (Afr)	–	–	*	*	*	–	–
South Africa (Eng)	–	–	*	*	*	–	–
Spain	–	–	0	0	–	–	–
Sweden	–	–	0	0	0	0	0
Switzerland (Fr)	–	–	0	0	0	0	0
Switzerland (Ger)	–	–	0	0	0	0	0
Switzerland (It)	–	–	0	0	0	0	0
Thailand	0	0	1	1	–	–	–
United States	0	0	0	0	0	0	0

* Final test booklets were unavailable for review.

– Did not participate.

1.7 STATISTICAL AND CONTENT-RELATED CONSIDERATIONS

TIMSS also conducted a set of elaborate statistical checks on the data to further ensure that items were performing comparably across countries. Although only a small number of items were found to be inappropriate for international comparisons, throughout the series of item-checking steps a number of reasons were discovered for differences in items across countries. Most of these were inadvertent changes in the items during the printing process, including omitting an item option or misprinting the graphics associated with an item. However, differences attributable to translation problems were found for an item or two in several countries.

Each country was provided with its item analysis information. Specially produced by the IEA Data Processing Center (DPC), these data contained automatic flags for a number of conditions that can indicate an item may not be performing properly. In addition, the Australian Council for Educational Research (ACER) produced graphical representations of item statistics for each participating country. Two countries deleted one Population 2 item each based on this information. Examples of these materials are displayed in Appendix A.

Prior to the international scaling of the Population 2 achievement data by ACER, the International Study Center conducted a thorough review of the item statistics for all participating countries. The process was empirically based, with data about the items being produced from several sources, including the translation verification process, the IEA DPC, and the item analysis information specially prepared for TIMSS by ACER. The intention was to detect inadvertent errors that were made during the processes of translation, printing, coding, and data entry.

As shown in Figure A.1 in Appendix A, the IEA DPC summarized on a page the item analysis results across countries for each item. The IEA DPC also produced information about the inter-rater agreement for the free-response items. ACER produced across-country graphical representations of item statistics, indices of fit, and item-by-country interactions. Figure A.2 in Appendix A provides an example of this type of item information. ACER screened the item statistics for particular problems, such as positive point-biserials for any non-key options and negative point-biserials for the key (see Figure A.3 in Appendix A). These summaries were particularly useful in highlighting items with potential problems.

In particular, items with the following problems were checked for possible deletion from the international database:

1. Errors were detected in the translation verification process that were not corrected.
2. The data cleaning process revealed more or fewer options than in the original version of the item.
3. The item analysis information showed the item to have a negative biserial.

4. The item-by-country interaction results showed a very large negative interaction for that country.
5. The item-fit statistic indicated the item was not fitting the model.
6. For free-response items, the within-country scoring reliability data showed an agreement of less than 70% for the score level. Also, performance in items with more than one score level was not ordered by score, or correct levels were associated with negative point-biserials.

The statistics and translation verification documentation were used as pointers toward checking actual booklets and contacting NRCs. If a problem could be detected by the International Study Center (such as a negative point-biserial for a correct answer or too few options for the multiple-choice questions), a decision was made to delete the item from the international scaling. However, if there was a question about potential translation or cultural issues, then the NRC was queried, and the International Study Center abided by the decision made by the NRC. In several cases, NRCs consulted mathematics or science experts before making a decision.

Considering that the checking involved approximately 500 items for each of more than 40 countries, very few deviations from the international format were revealed. Appendix B contains a list of the changes made in the international database for Population 2. Twenty-one countries had one or more items (usually only one) deleted as a result of translation, adaptation, or printing deviations. For three countries, options became switched in printing but were corrected in the database.

1.8 SUMMARY

Because the international versions of the TIMSS instruments were prepared in English, translating the materials into the 30 different languages of testing used by the 45 participating countries, and adapting the originals for countries testing in English, was an enormous challenge. Considerable energy, time, and resources were expended to help ensure that the translations yielded comparable test instruments across the countries.

The international English versions of the instruments were produced centrally by the TIMSS International Study Center, and then translated by the national centers in accordance with detailed translation guidelines. These guidelines included information about cultural adaptations, item difficulty, reading level, and layout as well as about retaining the meaning of the items. The specified procedures called for using two independent translators each for mathematics and science and having a third translator compare the versions. The best version was selected when there were differences in the translations. Most countries were able to adopt these procedures although resources were a difficulty in some instances.

The national centers submitted their translated instruments to the TIMSS International Study Center for review by professional translators. The translators completed verification forms noting any errors that needed to be corrected and making suggestions for improvements. These forms were returned to the national centers to make any necessary changes. Although time constraints made complete checking of camera-ready materials for all countries prior to testing impractical, the printed materials were checked thoroughly by both the TIMSS quality control monitors and the TIMSS International Study Center. Using information from the translation verification process in conjunction with a rigorous review of the item statistics did reveal some mistranslated and misprinted items for some countries. Such items are being deleted from the international database before scaling and analysis. Considering, however, the number of countries, languages, and items involved in the TIMSS testing, very few deviations from the international format were revealed in the final printed instruments. For example, at Population 2, most countries did not have any items deleted as a result of translation or printing problems.

REFERENCES

Hambleton, R.K. (1992). *Translating Achievement Tests for Use in Cross-National Studies* (Doc. Ref.: ICC454/NRC127). Paper prepared for the Third International Mathematics and Science Study.

Maxwell, B. (1996). "Translation and Cultural Adaptation of the Survey Instruments" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994a). *Survey Operations Manual– Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

2. SAMPLING

Pierre Foy
Michael O. Martin
Dana L. Kelly

2.1 OVERVIEW

The selection of valid and efficient samples is crucial to the quality and success of an international comparative study of student achievement. The accuracy of the survey results depends on the quality of the sampling information available, and particularly on the quality of the sampling itself. The procedures must therefore be explicit and practical and all steps must be documented fully. In a study as ambitious as TIMSS, the sample design and sampling procedures are complex, and the gathering of the required information about the national education systems places considerable demands on resources and expertise. Simplifying the sampling procedures to the extent possible, especially the sample selection within schools, was thus a major consideration in developing the sample design.

The sample design for TIMSS is described in detail in Foy, Rust, and Schleicher (1996). The basic design for Populations 1 (the pair of adjacent grades containing most 9-year-olds) and 2 (the pair of adjacent grades containing most 13-year-olds) consisted of a two-stage stratified probability sample of students, with schools stratified by nationally relevant variables and sampled with probability-proportional-to-size at the first stage, and a single intact class of students sampled from each of the two adjacent grade levels at the second stage. Countries were expected to sample at least 150 schools, although some of the larger countries chose to sample more, and some others were able to achieve satisfactory precision with less. This design was expected to yield a representative sample of approximately 7,500 students per country, with approximately 3,750 students at each grade level. NRCs were allowed to adapt the TIMSS sample design for their educational system, using more sampling information and more sophisticated sample designs and procedures than the base design provided. However, these solutions had to be approved and monitored by the international project management (the International Coordinating Center at the University of British Columbia until August 1993, and the International Study Center at Boston College thereafter).

The international project management provided various resources in the form of manuals, software programs, training, and continuous support to help NRCs identify a sample design appropriate for their national system, and to guide them through the phases of sampling. The *Sampling Plan* (TIMSS, 1992) provided an overview of the sample design and described the survey design options offered. The *Sampling Manual*

(TIMSS, 1994b) described how to implement the sampling plan and offered advice on initial planning, working within constraints, establishing appropriate sample selection procedures, and fieldwork. It provided an operational definition of the school sample and detailed the procedures for selecting it for Populations 1 and 2. The *Population 3 Sampling Guide* (TIMSS, 1994a) outlined the school sampling procedures for Population 3.

Included in the *Sampling Manual* (TIMSS, 1994b) were a number of forms that ensured that vital information at key stages was collected and recorded in a uniform manner for each country. Target population definitions, choice of stratifying variables, construction of school sampling frames, selection of school sample, and the like were therefore clearly documented. These forms were completed by NRCs and submitted to Statistics Canada for review and archiving. They are described in section 2.2 and displayed in Appendix C.

The *Survey Operations Manuals* (TIMSS, 1994e, 1994f) and *School Coordinators Manuals* (TIMSS, 1994c, 1994d) provided information on sampling of students within schools, the assignment of test booklets to sampled students, and administration and monitoring procedures used to identify and track respondents and nonrespondents. NRCs also received software designed to automate the sometimes complex within-school sampling procedures. This software was developed specially for TIMSS by the IEA Data Processing Center and Statistics Canada.

NRCs had several sources of expert support throughout all phases of sampling. Statistics Canada provided advice and support throughout the process. NRCs met with Statistics Canada staff during the semi-annual meetings of the National Research Coordinators and communicated regularly via fax, telephone, and e-mail. During consultation sessions, NRCs received training in how to select the school and student samples and in the use of the sampling software. In consultation with the TIMSS sampling referee (Keith Rust, WESTAT, Inc.) and the TIMSS Technical Advisory Committee, Statistics Canada reviewed the national sampling plans, sampling data, sampling frames, and sampling operations.

2.2 DOCUMENTATION OF THE SAMPLING PROCEDURES

NRCs were required to submit their completed sampling forms, described below, as documentation of the steps completed and of the quality of their samples. Information collected through these forms was used to evaluate the quality of the national samples and to categorize and annotate countries in the international reports. The required sampling forms related to three different aspects of the sampling process: population definition, sample design, and sample execution.

Statistics Canada was responsible for monitoring the sampling activities in the participating countries and for ensuring that all necessary documentation was received. Based on this documentation, the status of the national samples could be evaluated by the TIMSS sampling referee, the Technical Advisory Committee, and the International Study Center.

2.2.1 POPULATION DEFINITION

In order to obtain national samples for which to make meaningful comparisons, some initial steps needed to be completed and information provided. Forms 1 and 2 were used to document the required information. These forms were important since they would define the target population in terms of coverage, target grades, and exclusions. Although Form 1 was not a critical component, its contents did prove useful as scheduling and diagnostic tools. Compliance with reporting this information was very good. Table 2.1 shows for each country the status of the forms required to define the target population for Population 2.

Form 1 - TIMSS Participation and Primary School Structure

This form requested basic descriptive information on a national school system, namely the school calendar and expected testing dates, age-of-entry requirements, and grade structure through primary and secondary schooling. Although none of this was critical to the successful implementation of the sampling procedures, the forms did nonetheless provide useful information for determining field schedules and validating, to some degree, the target grades selected for TIMSS. Compliance with the delivery of this information was generally good.

Form 2/ Part 1 - Describing the National Desired Population

This form requested information used to establish the coverage of the definition of the national population and the target grades for TIMSS. This was very important, since selecting suitable grades was vital to the successful implementation of sampling procedures. Also, population coverage is an important piece of information to be reported for national school systems. Compliance with the delivery of these data was very good.

Form 2/Part 2 - Describing the National Defined Population

This form sought to identify all elements of the population that were to be excluded from the sampling process, at the school level as well as within schools. An important quality criterion for TIMSS was to limit all exclusions to less than 10% of the defined national coverage. Reporting this information was therefore important and compliance was very good.

Table 2.1
Status of Population Definition Documentation – Population 2

Country	Form 1	Form 2/P 1	Form 2/P 2	Notes
Argentina	C	C	C	Population coverage less than 100%
Australia	C	C	C	Target grades vary by state
Austria	C	C	C	
Belgium (Fl)	C	C	C	
Belgium (Fr)	C	C	C	
Bulgaria	C	C	C	
Canada	C	C	C	
Colombia	C	C	C	Students in selected grades older than expected
Cyprus	C	C	C	
Czech Republic	C	C	C	
Denmark	C	C	C	
England	C	C	C	Exclusion rate greater than 10%
France	C	C	C	
Germany	I	C	C	Population coverage less than 100%
Greece	C	C	C	
Hong Kong	P	C	C	
Hungary	I	P	C	
Iceland	I	P	P	
Indonesia	C	C	C	Population coverage less than 100%
Iran	C	C	C	
Ireland	C	C	C	
Israel	C	C	C	Population coverage less than 100% & only one target grade selected
Japan	C	C	C	
Korea	C	C	C	
Kuwait	I	C	C	Students in lone selected grade are older than expected
Latvia	C	C	C	Population coverage less than 100%
Lithuania	C	C	C	Population coverage less than 100%
Mexico	C	C	C	
Netherlands	C	C	C	
New Zealand	C	C	C	
Norway	C	C	C	
Philippines	C	C	P	Population coverage less than 100%
Portugal	C	C	C	
Romania	C	C	C	Students in selected grades older than expected
Russian Federation	C	C	C	
Scotland	C	C	C	
Singapore	C	C	C	
Slovak Republic	C	C	C	
Slovenia	I	C	C	Students in selected grades older than expected
South Africa	C	C	C	
Spain	C	C	C	
Sweden	C	C	C	
Switzerland	I	C	C	Population coverage less than 100%
Thailand	C	P	P	
United States	C	C	C	

C Complete information provided
P Partial information provided - adequate for monitoring
I Incomplete or no information provided

2.2.2 SAMPLE DESIGN

A number of forms of varying importance were used to document the national sample design for each country. The main purpose of these forms was to monitor the development of the sample designs. The importance of these forms varied depending on the complexity of the proposed sample designs. Compliance with reporting also varied, usually as a function of the complexity of the sample designs. As a general rule, countries that complied were successful in implementing their sample designs. Conversely, countries that had some difficulties in implementing their sample designs did not fully comply with the reporting requirements. Table 2.2 shows the status of the forms required for the sample design for Population 2.

Form 3 - Stratification Variables

On Form 3 NRCs were to report all variables used to stratify the school sampling frame. This information was not essential to the successful implementation of the sampling procedures, but advance knowledge was useful as a diagnostic tool to assist NRCs in developing their sample designs. Compliance with reporting was generally good, but this information could also be derived from the school sampling frames.

Form 4/Part 1 - Sample Design Structure

This form requested specific sample design details that would permit an evaluation of the adequacy of the planned sample size. Compliance with reporting this information was generally good but the quality was not always adequate. The quality was greatly improved through follow-up meetings with NRCs.

Form 4/Part 2 - Type of Sampling Frame

This form requested a description of the available school sampling frames. The form was not essential but proved useful to identify difficulties in finding adequate sampling frames and perhaps the need for more complex sample designs. Compliance with the delivery of this information was generally good.

Form 5/Part 1 - Schools Excluded From Sampling Frame

This form requested the list of all schools excluded from the school sampling frame. Compliance with the delivery of this information was not very good. However, given that Form 2 provided a description of all excluded schools as well as the number of students enrolled, the actual list of excluded schools was not considered an essential piece of information.

From 5/Part 2 - Recording the Formation of Pseudo-Schools

This form requested information on the construction of pseudo-schools. All countries that constructed pseudo-schools provided this form.

Form 6/Part 1 - Strata for Defined Population - Population Statistics

This form requested basic population counts of schools and students by strata for monitoring purposes. Although these data were not essential to the successful implementation of the sampling procedures, compliance with reporting was generally good.

Form 6/Part 2 - Strata for Defined Population - Sample Statistics

This form requested similar basic counts of schools and students, by strata, from the sample. Again, this information was not essential to the successful implementation of sampling procedures, since it could ultimately be derived from the data. It did nonetheless provide some indication of the total sample sizes. Compliance with the delivery of this information was generally good.

Table 2.2
Status of Sample Design Documentation – Population 2

Country	Form 3	Form 4/P 1	Form 4/P 2	Form 5/P 1	Form 5/P 2	Form 6/P 1	Form 6/P 2	Comments
Argentina	C	C	C	I	–	C	C	
Australia	C	C	C	I	–	C	C	
Austria	C	C	C	I	–	C	C	Sampled science classrooms
Belgium (Fl)	C	C	C	I	–	C	C	School subsample for upper grade vocational track
Belgium (Fr)	C	C	C	I	–	C	C	School subsample for upper grade vocational track
Bulgaria	C	C	C	C	C	C	C	
Canada	C	C	C	P	–	C	C	
Colombia	C	C	C	C	C	C	C	
Cyprus	C	C	C	C	C	C	C	All schools in sample
Czech Republic	C	C	C	C	C	C	C	
Denmark	C	C	C	I	–	C	C	Stratified SRS for schools (equal probabilities)
England	C	C	C	C	–	C	C	Sample of students, rather than classrooms
France	C	C	C	I	–	P	P	
Germany	P	P	P	I	–	C	C	Upper grade classrooms sampled with PPS
Greece	I	I	I	I	–	I	P	
Hong Kong	C	C	C	I	–	C	C	
Hungary	I	I	C	C	–	C	C	Classrooms sampled with PPS
Iceland	I	I	I	P	–	P	P	All schools in sample
Indonesia	C	C	C	C	C	C	C	
Iran	C	C	C	I	–	C	P	
Ireland	C	C	C	C	C	C	C	
Israel	C	C	C	C	C	C	C	
Japan	C	C	C	C	C	C	C	Stratified SRS for schools (equal probabilities)
Korea	C	C	C	C	C	C	C	
Kuwait	C	C	C	I	–	C	C	All schools in sample
Latvia	C	C	C	I	C	C	P	
Lithuania	C	C	C	I	–	P	P	
Mexico	C	C	C	P	–	C	C	
Netherlands	C	C	C	I	–	I	P	
New Zealand	C	C	C	C	C	C	C	
Norway	C	C	C	I	–	P	P	
Philippines	C	I	I	I	–	P	P	
Portugal	C	C	C	C	C	C	C	
Romania	C	C	C	C	C	C	C	
Russian Federation	C	C	C	P	P	P	P	Preliminary sampling stage
Scotland	C	C	C	C	C	C	C	
Singapore	C	C	C	C	C	C	C	All schools in sample
Slovak Republic	C	C	C	P	–	P	P	
Slovenia	C	I	I	I	–	C	C	
South Africa	C	C	C	I	–	C	C	
Spain	C	C	C	C	C	C	C	
Sweden	C	C	C	I	–	I	P	
Switzerland	C	C	C	I	–	C	C	
Thailand	I	I	I	I	–	I	P	Stratified SRS for schools (equal probabilities)
United States	C	C	C	I	–	P	P	Preliminary sampling stage

C Complete information provided
P Partial information provided - adequate for monitoring
I Incomplete or no information provided
– Not applicable

2.2.3 SAMPLE EXECUTION

The forms used to document the sample execution were very important since they demonstrated its success. Compliance with reporting this information was very good and generally indicative of the quality of the sample execution. Delivery of Forms 8 and 9 was not critical since the same information could be retrieved from Form 7 and the actual data files. Table 2.3 shows the status of the forms required for the sample execution. This table also presents additional comments for some countries, related to the information provided on the forms.

Form 7 - Sampling Frame and Sample Selection

This form requested the full school sampling frame with the sampled schools identified. This was important to validate the school sampling process. Compliance with the delivery of this information was very good, with a few notable exceptions. Argentina did not deliver its sampling frame and this was indicative of a major problem with the school sample. Eventually, lack of resources caused Argentina to discontinue participation in the study. The sampling frame provided by the Philippines was not documented in a way that supported the computation of satisfactory sampling weights. Selected unweighted results for the Philippines were presented in an appendix to the international reports. Germany also did not deliver its school sampling frame, but all other documentation indicates strongly that this school sample was selected properly. Indonesia, Scotland, and the United States delivered only partial school sampling frames, but enough to verify that the school samples were selected properly.

Form 8 - Identifying the Sample of Schools - Selection Numbers

This form requested the list of all random numbers used to select the sampled schools. This information was not essential since it could generally be derived from Form 7. Compliance with reporting was very good. The handful of countries that used alternate school sampling methods generally could not provide a corresponding Form 8, but their school sampling frames and other supporting documentation were sufficient to validate their school samples.

Form 9 - School Tracking Form

This form requested the list of all sampled schools along with their assigned replacements. It also indicated the participation status of sampled schools. Compliance with the delivery of this information was very good. The information could also be derived from the data.

Class Tracking Form

This form requested information on the classroom sampling procedures. This was a critical piece of information and compliance with its delivery was very good. Some countries that did not deliver this form were able to provide sufficient information to compute sampling weights. In other cases, the inability to deliver the form was indicative of problems in sampling classrooms. This was the case for Denmark, Greece, and Thailand. Some countries that delivered only partial forms provided additional or alternate documentation, usually in the form of spreadsheets, to describe this aspect of the sampling process.

Table 2.3
Status of Sampling Execution Documentation – Population 2

Country	Form 7	Form 8	Form 9	CTF	Comments
Argentina	I	I	C	C	Unapproved school sampling procedure
Australia	C	C	C	C	
Austria	C	C	C	C	
Belgium (Fl)	C	C	C	C	
Belgium (Fr)	C	C	C	C	
Bulgaria	C	P	P	C	
Canada	C	C	C	C	
Colombia	C	C	C	C	
Cyprus	C	–	C	C	
Czech Republic	C	C	C	C	
Denmark	C	–	P	P	Unapproved classroom sampling procedure
England	C	C	P	–	
France	C	–	C	C	
Germany	I	C	C	P	School sampling frame not available
Greece	C	P	C	P	Unapprovided classroom sampling procedure
Hong Kong	C	P	C	C	
Hungary	C	C	C	P	Classroom selection probabilities not always correct
Iceland	P	–	P	C	
Indonesia	I	C	C	C	
Iran	C	C	C	C	
Ireland	C	C	C	C	
Israel	C	C	C	C	
Japan	C	–	C	C	
Korea	C	C	C	C	
Kuwait	C	–	P	C	Unapproved classroom sampling procedure
Latvia	C	C	C	P	
Lithuania	C	P	P	P	
Mexico	C	C	C	C	
Netherlands	C	C	C	C	
New Zealand	C	C	C	C	
Norway	C	P	P	C	
Philippines	C	I	P	C	Documentation inadequate to compute sampling weights
Portugal	C	C	C	C	
Romania	C	C	C	C	
Russian Federation	C	C	C	C	
Scotland	I	P	C	C	
Singapore	C	–	C	C	
Slovak Republic	C	C	C	C	
Slovenia	C	C	P	C	
South Africa	C	C	C	C	Non-participating students not recorded
Spain	C	C	C	C	
Sweden	C	–	P	C	
Switzerland	C	C	C	P	
Thailand	C	–	P	P	Unapproved classroom sampling procedure
United States	I	P	C	C	

C Complete information provided
P Partial information provided - adequate for monitoring
I Incomplete or no information provided
– Not applicable

2.3 POPULATION DEFINITIONS AND SAMPLE PARTICIPATION RATES

Tables 2.4 through 2.11 summarize the status of the TIMSS Population 2 samples as of September 25, 1996.

Table 2.4 describes the coverage of the population definitions in each country. In IEA studies, the *International Desired Population* is the population for which, ideally, results are required. For Population 2 in TIMSS, the international desired population consisted of all students in the country who were enrolled in one of the two adjacent grades containing the highest proportion of students aged 13 years at the time of testing. The *National Desired Population* for a country should correspond closely to this, and its coverage of the international desired population should ideally be 100%. In cases where it was not possible to implement the international desired population without modification, TIMSS permitted a country to define a national desired population that did not include part of the international desired population. Where this occurred it was the result of the exclusion of certain geographic or political units, language groups, or distinct school system components. The first column of figures in Table 2.4 gives the percentage coverage for each of the TIMSS participants. Just eight of the participants reported coverage less than 100%, and these are annotated in the international reports.

The *National Defined Population* consists of that portion of the country's national desired population that was covered by the school, classroom, and student sampling procedures and thus had a chance of being selected in the country's sample of students. Differences between the national desired populations and national defined populations could result from excluding schools (e.g., very small schools, or schools in remote areas), and from excluding certain kinds of students (e.g., students with physical and learning disabilities who were unable to take the assessment under TIMSS testing conditions). The remaining columns in Table 2.4 contain the percentages of the national desired population that were excluded by each participant. Countries where the overall exclusions exceed 10% are annotated in the international reports.

The two adjacent grades that the contained most 13-year-olds were the seventh and eighth grades in many countries. Table 2.5 records the grades tested in each country, with the names for those grades as provided by the participants. Table 2.6 presents the percentage of 13-year-olds in the grades tested in each country. The achievement results for countries not testing the two grades containing the most 13-year-olds are presented in a separate section of tables in the international reports.

Table 2.7 presents school participation rates and sample sizes for the eighth-grade sample. The table includes the weighted school participation rate before and after replacement of non-participating schools, the number of schools in the originally selected sample, the number of these schools that were in fact eligible for selection, the number of schools in the originally selected sample that participated, the number of replacement schools that participated, and the total number of participating schools.

Table 2.8 presents student participation rates and sample sizes for the eighth-grade sample. The table includes the weighted student participation rate, the number of students in participating schools, the number of students withdrawn from sampled schools or classrooms before the test administration, the number of students excluded, the number of eligible and absent students in the sampled classrooms, and the total number of students assessed.

Tables 2.9 and 2.10 present the same information for the seventh-grade sample as Table 2.6 and 2.7 present for the eighth-grade.

Table 2.11 presents the overall weighted participation rates for the seventh-grade and eighth-grade samples both before and after the inclusion of replacement schools.

Table 2.4
Coverage of TIMSS Target Population

The International Desired Population is defined as follows:
Population 2 - All students enrolled in the two adjacent grades with the largest proportion of 13-year-old students at the time of testing.

Country	International Desired Population		National Desired Population		
	Coverage	Notes on Coverage	School-Level Exclusions	Within-Sample Exclusions	Overall Exclusions
Australia	100%		0.2%	0.7%	0.8%
Austria	100%		2.9%	0.2%	3.1%
Belgium (Fl)	100%		3.8%	0.0%	3.8%
Belgium (Fr)	100%		4.5%	0.0%	4.5%
Bulgaria	100%		0.6%	0.0%	0.6%
Canada	100%		2.4%	2.1%	4.5%
Colombia	100%		3.8%	0.0%	3.8%
Cyprus	100%		0.0%	0.0%	0.0%
Czech Republic	100%		4.9%	0.0%	4.9%
Denmark	100%		0.0%	0.0%	0.0%
[2] England	100%		8.4%	2.9%	11.3%
France	100%		2.0%	0.0%	2.0%
[1] Germany	88%	15 of 16 regions*	8.8%	0.9%	9.7%
Greece	100%		1.5%	1.3%	2.8%
Hong Kong	100%		2.0%	0.0%	2.0%
Hungary	100%		3.8%	0.0%	3.8%
Iceland	100%		1.7%	2.9%	4.5%
Iran, Islamic Rep.	100%		0.3%	0.0%	0.3%
Ireland	100%		0.0%	0.4%	0.4%
[1] Israel	74%	Hebrew Public Education System	3.1%	0.0%	3.1%
Japan	100%		0.6%	0.0%	0.6%
Korea	100%		2.2%	1.6%	3.8%
Kuwait	100%		0.0%	0.0%	0.0%
[1] Latvia (LSS)	51%	Latvian-speaking schools	2.9%	0.0%	2.9%
[1] Lithuania	84%	Lithuanian-speaking schools	6.6%	0.0%	6.6%
Netherlands	100%		1.2%	0.0%	1.2%
New Zealand	100%		1.3%	0.4%	1.7%
Norway	100%		0.3%	1.9%	2.2%
Philippines	91%	2 provinces and autonomous regions excluded	6.5%	0.0%	6.5%
Portugal	100%		0.0%	0.3%	0.3%
Romania	100%		2.8%	0.0%	2.8%
Russian Federation	100%		6.1%	0.2%	6.3%
Scotland	100%		0.3%	1.9%	2.2%
Singapore	100%		4.6%	0.0%	4.6%
Slovak Republic	100%		7.4%	0.1%	7.4%
Slovenia	100%		2.4%	0.2%	2.6%
South Africa	100%		9.6%	0.0%	9.6%
Spain	100%		6.0%	2.7%	8.7%
Sweden	100%		0.0%	0.9%	0.9%
[1] Switzerland	86%	22 of 26 cantons	4.4%	0.8%	5.3%
Thailand	100%		6.2%	0.0%	6.2%
United States	100%		0.4%	1.7%	2.1%

[1] National Desired Population does not cover all of International Desired Population. Because coverage falls below 65%, Latvia is annotated LSS for Latvian Speaking Schools only.

[2] National Defined Population covers less than 90 percent of National Desired Population.

* One region (Baden-Wuerttemberg) did not participate.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.5
Information About the Grades Tested

Country	Lower Grade		Upper Grade	
	Country's Name for Lower Grade	Years of Formal Schooling Including Lower Grade[1]	Country's Name for Upper Grade	Years of Formal Schooling Including Upper Grade[1]
[2] Australia	7 or 8	7 or 8	8 or 9	8 or 9
Austria	3. Klasse	7	4. Klasse	8
Belgium (Fl)	1A	7	2A & 2P	8
Belgium (Fr)	1A	7	2A & 2P	8
Bulgaria	7	7	8	8
Canada	7	7	8	8
Colombia	7	7	8	8
Cyprus	7	7	8	8
Czech Republic	7	7	8	8
Denmark	6	6	7	7
England	Year 8	8	Year 9	9
France	5ème	7	4ème (90%) or 4ème Technologique (10%)	8
Germany	7	7	8	8
Greece	Secondary 1	7	Secondary 2	8
Hong Kong	Secondary 1	7	Secondary 2	8
Hungary	7	7	8	8
Iceland	7	7	8	8
Iran, Islamic Rep.	7	7	8	8
Ireland	1st Year	7	2nd Year	8
Israel	–	–	8	8
Japan	1st Grade Lower Secondary	7	2nd Grade Lower Secondary	8
Korea, Republic of	1st Grade Middle School	7	2nd Grade Middle School	8
Kuwait	–	–	9	9
Latvia	7	7	8	8
Lithuania	7	7	8	8
Netherlands	Secondary 1	7	Secondary 2	8
[3,4] New Zealand	Form 2	7.5 - 8.5	Form 3	8.5 - 9.5
[3] Norway	6	6	7	7
[3] Philippines	Grade 6 Elementary	6	1st Year High School	7
Portugal	Grade 7	7	Grade 8	8
Romania	7	7	8	8
[5] Russian Federation	7	6 or 7	8	7 or 8
Scotland	Secondary 1	8	Secondary 2	9
Singapore	Secondary 1	7	Secondary 2	8
Slovak Republic	7	7	8	8
Slovenia	7	7	8	8
Spain	7 EGB	7	8 EGB	8
[3] South Africa	Standard 5	7	Standard 6	8
[3] Sweden	6	6	7	7
[3] Switzerland				
(German)	6	6	7	7
(French and Italian)	7	7	8	8
Thailand	Secondary 1	7	Secondary 2	8
United States	7	7	8	8

[1] Years of schooling based on the number of years children in the grade level have been in formal schooling, beginning with primary education (International Standard Classification of Education Level 1). Does not include preprimary education.

[2] Australia: Each state/territory has its own policy regarding age of entry to primary school. In 4 of the 8 states/territories students were sampled from grades 7 and 8; in the other four states/territories students were sampled from grades 8 and 9.

[3] Indicates that there is a system-split between the lower and upper grades. In Switzerland there is a system-split in 14 of 26 cantons.

[4] New Zealand: The majority of students begin primary school on or near their 5th birthday so the "years of formal schooling" vary.

[5] Russian Federation: 70% of students in the seventh grade have had 6 years of formal schooling; 70% in the eighth grade have had 7 years of formal schooling.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95. Information provided by TIMSS National Research Coordinators

Table 2.6
Coverage of 13-Year-Old Students

Country	Percent of 13-Year-Olds in Lower Grade (Seventh Grade*)	Percent of 13-Year-Olds in Upper Grade (Eighth Grade*)	Percent of 13-Year-Olds in Both Grades
Australia	64%	28%	92%
Austria	62%	27%	89%
Belgium (Fl)	46%	49%	94%
Belgium (Fr)	41%	46%	87%
Bulgaria	58%	37%	95%
Canada	48%	43%	91%
Colombia	30%	15%	45%
Cyprus	28%	70%	98%
Czech Republic	73%	17%	90%
Denmark	35%	64%	98%
England	57%	42%	99%
France	44%	35%	78%
Germany	71%	2%	73%
Greece	11%	85%	96%
Hong Kong	44%	46%	90%
Hungary	65%	24%	89%
Iceland	16%	83%	100%
Iran, Islamic Rep.	47%	25%	72%
Ireland	69%	17%	86%
Israel	–	–	–
Japan	91%	9%	100%
Korea	70%	28%	98%
Kuwait	–	–	–
Latvia (LSS)	60%	26%	86%
Lithuania	64%	26%	90%
Netherlands	59%	31%	90%
New Zealand	52%	47%	99%
Norway	43%	57%	100%
Philippines	–	–	–
Portugal	44%	32%	76%
Romania	67%	9%	76%
Russian Federation	50%	44%	95%
Scotland	24%	75%	99%
Singapore	82%	15%	97%
Slovak Republic	73%	22%	95%
Slovenia	65%	2%	67%
South Africa	36%	20%	55%
Spain	46%	39%	85%
Sweden	45%	54%	99%
Switzerland	48%	44%	92%
Thailand	58%	20%	78%
United States	58%	33%	91%

*Seventh and eighth grades in most countries; see Table 2.5 for more information about the grades tested in each country. A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower (seventh) grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.7
School Participation Rates and Sample Sizes
Upper Grade (Eighth Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75%	77%	214	214	158	3	161
Austria	41%	84%	159	159	62	62	124
Belgium (Fl)	61%	94%	150	150	92	49	141
Belgium (Fr)	57%	79%	150	150	85	34	119
Bulgaria	72%	74%	167	167	111	4	115
Canada	90%	91%	413	388	363	1	364
Colombia	91%	93%	150	150	136	4	140
Cyprus	100%	100%	55	55	55	0	55
Czech Republic	96%	100%	150	149	143	6	149
Denmark	93%	93%	158	157	144	0	144
England	56%	85%	150	144	80	41	121
France	86%	86%	151	151	127	0	127
Germany	72%	93%	153	150	102	32	134
Greece	87%	87%	180	180	156	0	156
Hong Kong	82%	82%	105	104	85	0	85
Hungary	100%	100%	150	150	150	0	150
Iceland	98%	98%	161	132	129	0	129
Iran, Islamic Rep.	100%	100%	192	191	191	0	191
Ireland	84%	89%	150	149	125	7	132
Israel	45%	46%	100	100	45	1	46
Japan	92%	95%	158	158	146	5	151
Korea	100%	100%	150	150	150	0	150
Kuwait	100%	100%	69	69	69	0	69
Latvia (LSS)	83%	83%	170	169	140	1	141
Lithuania	96%	96%	151	151	145	0	145
Netherlands	24%	63%	150	150	36	59	95
New Zealand	91%	99%	150	150	137	12	149
Norway	91%	97%	150	150	136	10	146
Philippines	96% **	97% **	200	200	192	1	193
Portugal	95%	95%	150	150	142	0	142
Romania	94%	94%	176	176	163	0	163
Russian Federation	97%	100%	175	175	170	4	174
Scotland	79%	83%	153	153	119	8	127
Singapore	100%	100%	137	137	137	0	137
Slovak Republic	91%	97%	150	150	136	9	145
Slovenia	81%	81%	150	150	121	0	121
South Africa	60%	64%	180	180	107	7	114
Spain	96%	100%	155	154	147	6	153
Sweden	97%	97%	120	120	116	0	116
Switzerland	93%	95%	259	258	247	3	250
Thailand	99%	99%	150	150	147	0	147
United States	77%	85%	220	217	169	14	183

*Eighth grade in most countries; see Table 2.5 for more information about the grades tested in each country.

**Participation rates for the Philippines are unweighted.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.8

Student Participation Rates and Sample Sizes Upper Grade (Eighth Grade*)

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	92%	8027	63	61	7903	650	7253
Austria	95%	2969	14	4	2951	178	2773
Belgium (Fl)	97%	2979	1	0	2978	84	2894
Belgium (Fr)	91%	2824	0	1	2823	232	2591
Bulgaria	86%	2300	0	0	2300	327	1973
Canada	93%	9240	134	206	8900	538	8362
Colombia	94%	2843	6	0	2837	188	2649
Cyprus	97%	3045	15	0	3030	107	2923
Czech Republic	92%	3608	6	0	3602	275	3327
Denmark	93%	2487	0	0	2487	190	2297
England	91%	2015	37	60	1918	142	1776
France	95%	3141	0	0	3141	143	2998
Germany	87%	3318	0	35	3283	413	2870
Greece	97%	4154	27	23	4104	114	3990
Hong Kong	98%	3415	12	0	3403	64	3339
Hungary	87%	3339	0	0	3339	427	2912
Iceland	90%	2025	10	65	1950	177	1773
Iran, Islamic Rep.	98%	3770	20	0	3750	56	3694
Ireland	91%	3411	28	10	3373	297	3076
Israel	98%	1453	6	0	1447	32	1415
Japan	95%	5441	0	0	5441	300	5141
Korea	95%	2998	31	0	2967	47	2920
Kuwait	83%	1980	3	0	1977	322	1655
Latvia (LSS)	90%	2705	19	0	2686	277	2409
Lithuania	87%	2915	2	0	2913	388	2525
Netherlands	95%	2112	14	1	2097	110	1987
New Zealand	94%	4038	121	12	3905	222	3683
Norway	96%	3482	26	49	3407	140	3267
Philippines	91% **	6586	93	0	6493	492	6001
Portugal	97%	3589	70	13	3506	115	3391
Romania	96%	3899	0	0	3899	174	3725
Russian Federation	95%	4311	42	10	4259	237	4022
Scotland	88%	3289	0	46	3243	380	2863
Singapore	95%	4910	18	0	4892	248	4644
Slovak Republic	95%	3718	5	3	3710	209	3501
Slovenia	95%	2869	15	8	2846	138	2708
South Africa	97%	4793	0	0	4793	302	4491
Spain	95%	4198	27	102	4069	214	3855
Sweden	93%	4483	71	28	4384	309	4075
Switzerland	98%	4989	16	24	4949	94	4855
Thailand	100%	5850	0	0	5850	0	5850
United States	92%	8026	104	108	7814	727	7087

*Eighth grade in most countries; see Table 2.5 for more information about the grades tested in each country.
**Participation rates for the Philippines are unweighted.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.9

School Participation Rates and Sample Sizes Lower Grade (Seventh Grade*)

Country	School Participation Before Replacement (Weighted Percentage)	School Participation After Replacement (Weighted Percentage)	Number of Schools in Original Sample	Number of Eligible Schools in Original Sample	Number of Schools in Original Sample That Participated	Number of Replacement Schools That Participated	Total Number of Schools That Participated
Australia	75%	76%	214	213	156	3	159
Austria	43%	86%	159	159	63	62	125
Belgium (Fl)	61%	93%	150	150	91	49	140
Belgium (Fr)	57%	80%	150	150	85	35	120
Bulgaria	75%	77%	150	150	101	3	104
Canada	90%	90%	413	390	366	1	367
Colombia	91%	93%	150	150	136	4	140
Cyprus	100%	100%	55	55	55	0	55
Czech Republic	96%	100%	150	150	144	6	150
Denmark	88%	88%	158	154	137	0	137
England	57%	85%	150	145	81	41	122
France	87%	87%	151	151	126	0	126
Germany	70%	90%	153	153	101	31	132
Greece	87%	87%	180	180	156	0	156
Hong Kong	83%	83%	105	104	86	0	86
Hungary	99%	99%	150	150	149	0	149
Iceland	97%	97%	161	149	144	0	144
Iran, Islamic Rep.	100%	100%	192	192	192	0	192
Ireland	82%	87%	150	148	122	7	129
Israel	–	–	–	–	–	–	–
Japan	92%	95%	158	158	146	5	151
Korea	100%	100%	150	150	150	0	150
Kuwait	–	–	–	–	–	–	–
Latvia (LSS)	83%	84%	170	169	141	1	142
Lithuania	96%	96%	151	151	145	0	145
Netherlands	23%	61%	150	150	34	58	92
New Zealand	90%	99%	150	150	135	13	148
Norway	84%	96%	150	147	124	17	141
Philippines	97% **	97% **	200	200	194	0	194
Portugal	94%	94%	150	150	141	0	141
Romania	94%	94%	176	175	162	0	162
Russian Federation	97%	100%	175	175	170	4	174
Scotland	79%	85%	153	153	120	9	129
Singapore	100%	100%	137	137	137	0	137
Slovak Republic	91%	97%	150	150	136	9	145
Slovenia	81%	81%	150	150	122	0	122
South Africa	83%	85%	161	161	133	4	137
Spain	96%	100%	155	154	147	6	153
Sweden	96%	96%	160	160	154	0	154
Switzerland	90%	94%	217	217	200	6	206
Thailand	99%	99%	150	150	146	0	146
United States	77%	84%	220	214	165	14	179

*Seventh grade in most countries; see Table 2.5 for more information about the grades tested in each country.

**Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.10

Student Participation Rates and Sample Sizes Lower Grade (Seventh Grade*)

Country	Within School Student Participation (Weighted Percentage)	Number of Sampled Students in Participating Schools	Number of Students Withdrawn from Class/School	Number of Students Excluded	Number of Students Eligible	Number of Students Absent	Total Number of Students Assessed
Australia	93%	6067	26	21	6020	421	5599
Austria	95%	3196	22	5	3169	156	3013
Belgium (Fl)	97%	2857	3	0	2854	86	2768
Belgium (Fr)	95%	2418	0	1	2417	125	2292
Bulgaria	87%	2080	0	0	2080	282	1798
Canada	95%	8962	89	248	8625	406	8219
Colombia	93%	2840	2	0	2838	183	2655
Cyprus	98%	3028	17	0	3011	82	2929
Czech Republic	92%	3641	11	0	3630	285	3345
Denmark	86%	2408	0	0	2408	335	2073
England	92%	2031	31	67	1933	130	1803
France	95%	3164	0	0	3164	148	3016
Germany	87%	3388	0	37	3351	458	2893
Greece	97%	4166	30	78	4058	127	3931
Hong Kong	98%	3507	11	0	3496	83	3413
Hungary	94%	3266	0	0	3266	200	3066
Iceland	92%	2243	11	72	2160	203	1957
Iran, Islamic Rep.	99%	3789	18	0	3771	36	3735
Ireland	91%	3480	23	17	3440	313	3127
Israel	–	–	–	–	–	–	–
Japan	96%	5337	0	0	5337	207	5130
Korea	94%	2996	51	0	2945	38	2907
Kuwait	–	–	–	–	–	–	–
Latvia (LSS)	91%	2853	7	0	2846	279	2567
Lithuania	89%	2852	3	0	2849	318	2531
Netherlands	95%	2220	23	0	2197	100	2097
New Zealand	95%	3471	98	17	3356	172	3184
Norway	96%	2629	8	53	2568	99	2469
Philippines	93% **	6283	29	1	6253	401	5852
Portugal	96%	3594	80	4	3510	148	3362
Romania	95%	3938	0	0	3938	192	3746
Russian Federation	96%	4408	39	11	4358	220	4138
Scotland	90%	3313	0	81	3232	319	2913
Singapore	98%	3744	19	0	3725	84	3641
Slovak Republic	95%	3797	10	3	3784	184	3600
Slovenia	95%	3058	12	4	3042	144	2898
South Africa	96%	5532	0	0	5532	231	5301
Spain	95%	4087	38	116	3933	192	3741
Sweden	95%	3055	27	36	2992	161	2831
Switzerland	99%	4199	14	44	4141	56	4085
Thailand	100%	5845	0	0	5845	0	5845
United States	94%	4295	42	85	4168	282	3886

*Seventh grade in most countries; see Table 2.5 for more information about the grades tested in each country.

**Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 2.11

Overall Participation Rates
Upper and Lower Grades (Seventh and Eighth Grades*)

Country	Upper Grade		Lower Grade	
	Overall Participation Before Replacement (Weighted Percentage)	Overall Participation After Replacement (Weighted Percentage)	Overall Participation Before Replacement (Weighted Percentage)	Overall Participation After Replacement (Weighted Percentage)
Australia	69%	70%	69%	71%
Austria	39%	80%	41%	82%
Belgium (Fl)	59%	91%	59%	91%
Belgium (Fr)	52%	72%	54%	76%
Bulgaria	62%	63%	65%	67%
Canada	84%	84%	86%	86%
Colombia	85%	87%	84%	86%
Cyprus	97%	97%	98%	98%
Czech Republic	89%	92%	88%	92%
Denmark	86%	86%	76%	76%
England	51%	77%	52%	78%
France	82%	82%	82%	82%
Germany	63%	81%	61%	78%
Greece	84%	84%	84%	84%
Hong Kong	81%	81%	81%	81%
Hungary	87%	87%	93%	93%
Iceland	88%	88%	89%	89%
Iran, Islamic Rep.	98%	98%	99%	99%
Ireland	76%	81%	75%	79%
Israel	44%	45%	–	–
Japan	87%	90%	88%	91%
Korea	96%	95%	94%	94%
Kuwait	83%	83%	–	–
Latvia (LSS)	75%	75%	75%	76%
Lithuania	83%	83%	86%	86%
Netherlands	23%	60%	22%	58%
New Zealand	86%	94%	85%	94%
Norway	87%	93%	81%	92%
Philippines	87% **	88% **	90% **	90% **
Portugal	92%	92%	90%	90%
Romania	89%	89%	89%	89%
Russian Federation	93%	95%	93%	95%
Scotland	69%	73%	71%	76%
Singapore	95%	95%	96%	98%
Slovak Republic	86%	91%	86%	92%
Slovenia	77%	77%	77%	77%
South Africa	58%	62%	79%	82%
Spain	91%	94%	91%	95%
Sweden	90%	90%	91%	91%
Switzerland	92%	94%	89%	93%
Thailand	99%	99%	99%	99%
United States	71%	78%	72%	79%

*Seventh and eighth grades in most countries; see Table 2.5 for information about the grades tested in each country.

** Participation rates for the Philippines are unweighted.

A dash (–) indicates data are unavailable. Israel and Kuwait did not test the lower grade.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

46

2.4 REPORTING ACHIEVEMENT RESULTS

The manner in which the achievement results for participants are presented in international reports was influenced by their sampling participation rates. Countries were assigned to one of three categories on the basis of their sampling participation.

Category 1 Acceptable sampling participation rate **without** the use of replacement schools.

Countries in this category will appear in the tables and figures in international reports without annotation, and will be ordered by achievement as appropriate.

Category 2 Acceptable sampling participation rate **only when replacement schools are included**.

Countries in this category will be annotated with a "dagger" in the tables and figures in international reports, and will be ordered by achievement as appropriate.

Category 3 Unacceptable sampling response rate even when replacement schools are included.

Countries in this category will appear in a separate section of the achievement tables, below the other countries, in international reports. These countries will be presented in alphabetical order.

In order to be placed in Category 1, a country had to have:

- An **unweighted** school response rate **without** replacement of at least 85% (after rounding to nearest whole percent) AND an **unweighted** student response rate (after rounding) of at least 85%

OR

- A **weighted** school response rate **without** replacement of at least 85% (after rounding to nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85%

OR

- The product of the (unrounded) **weighted** school response rate **without** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the nearest whole percent).

A country was placed in Category 2 if:

- It failed to meet the requirements for Category 1 but had a weighted school response rate **without** replacement of at least 50% (after rounding to the nearest percent)

AND EITHER

- A **weighted** school response rate **with** replacement of at least 85% (after rounding to nearest whole percent) AND a **weighted** student response rate (after rounding) of at least 85%

OR

- The product of the (unrounded) **weighted** school response rate **with** replacement and the (unrounded) **weighted** student response rate of at least 75% (after rounding to the nearest whole percent).

Countries that could provide documentation to show that they complied with TIMSS sampling procedures and requirements but did not meet the requirements for Category 1 or Category 2 were placed in Category 3.

2.5 SUMMARY

An enormous amount of time and effort was devoted to sampling issues and activities in TIMSS. The study is by far the largest comparative international survey of student achievement conducted to date, and by far the most demanding in terms of sampling requirements. The TIMSS data collection was conducted simultaneously in 45 countries, with three student populations incorporating five grade levels and two school subjects. In Population 2 alone, more than 300,000 students in more than 7,500 schools were sampled to take part in the study.

The study broke new ground, not only by the scale of its sampling operations and the care and attention that was paid to all aspects of the process, but also by the extent to which each stage of the procedure was documented and verified by the National Research Coordinators, the TIMSS sampling consultants, and the sampling referee. This emphasis on documentation was carried through to the reporting of results, where countries with irregularities in their sampling are clearly labeled, annotated, or presented in separate sections of tables, depending on the nature of the irregularity.

As documented in this report, the majority of participants in TIMSS did an excellent job in discharging their sampling responsibilities, and readers and reviewers of international reports may be assured that the results are based on accurate and well-documented samples. Perhaps inevitably for a cooperative venture on such a scale, there were some participants who found it difficult to complete all of their tasks in a satisfactory manner, but all such deficiencies are clearly labeled when data are reported, and should not be allowed to detract from the high professional standard achieved by most participants.

REFERENCES

Foy, P., Rust, K., and Schleicher, A. (1996). Sample Design in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1992). *Sampling Plan* (Doc. Ref.: ICC 438/NRC116). Prepared by Richard Wolfe and David Wiley.

Third International Mathematics and Science Study (TIMSS). (1994a). *Population 3 Sampling Guide* (Doc. Ref.: ICC917/NRC448). Prepared by Jean Dumais. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Sampling Manual, Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *School Coordinators Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *School Coordinators Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *Survey Operations Manual–Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994f). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

3. MONITORING THE TIMSS DATA COLLECTION

Michael O. Martin
Craig D. Hoyle
Kelvin D. Gregory

3.1 THE TIMSS QUALITY CONTROL MONITORS

Since all data collection activities took place within participating countries and were the responsibility of NRCs, it was considered essential to have a representative of the International Study Center visit each country to interview the NRC about data collection plans and procedures and to select and visit a sample of schools while the TIMSS testing was taking place.

In December 1994, the International Study Center contracted Goodison Associates (United States) to help with the hiring and training of a team of quality control monitors to carry out the required visits. Goodison Associates also helped to develop a procedural manual and data collection instruments for the quality control monitors, and were responsible for paying them and ensuring that they met their contractual obligations.

In January 1995, NRCs were asked to nominate a person, such as a retired school teacher, to serve as quality control monitor for their country. The International Study Center reviewed the nominations and in almost all cases adopted the NRC's first suggestion. The monitors were trained centrally before returning to their countries to interview the NRC and to observe classroom testing sessions.

3.2 TRAINING OF QUALITY CONTROL MONITORS

The TIMSS quality control monitors were trained in a two-day session during which they were briefed on the design and purpose of TIMSS, the responsibilities of the NRC in conducting the study in each country, and their own roles and responsibilities. In total, five such training sessions were held. Most quality control monitors were trained in one of three scheduled sessions: February 1995, London, England; March 1995, Enschede, The Netherlands; April 1995, Paris, France. Two additional sessions were held to train the remaining quality control monitors, from Argentina (August 1995, Philadelphia, United States) and Australia and New Zealand (July 1995, Wellington, New Zealand). The quality control monitors are listed in Appendix D. Also provided in Appendix D is information about the training sessions.

The *Manual for the TIMSS Quality Control Monitors* (TIMSS, 1995d) was developed by the TIMSS International Study Center with the assistance of Goodison Associates and was used as the basis for the training sessions. The manual included:

- An introduction to TIMSS, outlining the purpose of the study, study schedule, management arrangements, the major components of TIMSS (populations, sampling design, test and questionnaire design), and the purpose of the quality assurance program
- An overview of the roles and responsibilities of the TIMSS quality control monitor
- An overview of the major tasks of the NRC
- Instructions for visiting the national center, interviewing the NRC, collecting the required materials from the NRC, and using the translation verification report to check the implementation of the suggestions made in the international review of the translations
- A questionnaire to be completed during the interview with the NRC
- Step-by-step procedures for selecting the schools for classroom observation
- Instructions for visiting these schools: arranging the visit, observing the testing sessions, completing the Classroom Observation Record, and interviewing the School Coordinator
- A copy of the Classroom Observation Record
- Instructions for returning materials to the International Study Center.

In addition to the *Manual for TIMSS Quality Control Monitors* (TIMSS, 1995d), each quality control monitor received copies of the *Survey Operations Manuals* (TIMSS, 1994d, 1994e), the *Test Administrator Manual* (TIMSS, 1994f), the *School Coordinator Manuals* (TIMSS, 1994b, 1994c), and the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c), which describe the procedures required for the implementation of TIMSS in each country. Although quality control monitors did not need to know every TIMSS policy and procedure in detail, they were encouraged to read through all the manuals in order to become familiar with the work of NRCs and the procedures to be followed in each country participating in TIMSS.

During each training session, a TIMSS International Study Center staff member explained the structure and major components of the study, emphasizing the NRC's tasks, especially as they related to the quality control monitor's duties. Goodison Associates staff members reviewed the quality control monitors' roles and responsibilities and led them through the schedule for the Interview with the National Research Coordinator and the Classroom Observation Record. Quality control monitors also took part in an exercise to help them select the schools for classroom observation.

3.3 THE QUALITY CONTROL MONITOR'S VISIT TO THE NATIONAL CENTER

The quality control monitor in each country was required to visit the TIMSS national center to (1) interview the National Research Coordinator about aspects of the data collection activities, (2) work with the NRC to select a sample of schools to visit, and (3) collect copies of the national versions of the TIMSS data collection instruments.

The quality control monitor's interview with the National Research Coordinator addressed the NRC's ten major responsibilities:

- Selecting the sample of students to be tested
- Working with the School Coordinators
- Translating the test instruments
- Assembling and printing the test booklets
- Packing and shipping the necessary materials to the designated School Coordinators
- Arranging for the return of materials from the school sites
- Arranging for coding the free-response and performance assessment questions
- Entering the testing results and information from students, teachers, and principals
- Conducting on-site quality assurance observations for a 10% sample of schools
- Preparing a report on survey activities.

The quality control monitor recorded the NRC's responses to questions regarding the implementation of these responsibilities, and any additional comments made regarding the TIMSS procedures. The interview questions were designed to ascertain the degree to which the procedures and policies described in the *Survey Operations Manuals* (TIMSS, 1994d, 1994e) the *Sampling Manual* (TIMSS, 1994a), the *Guide to Coding, Checking, and Entering the TIMSS Data* (TIMSS, 1995c), and other documents were followed.

3.4 SUMMARY OF RESULTS OF INTERVIEWS WITH NRCS

This section summarizes the main issues arising from the interviews. The data are presented in summary form in Appendix E. As shown in Table 3.1, interviews were conducted in all but four countries.

3.4.1 SAMPLING PROCEDURES

Most NRCs reported that they were able to select a sample of schools and students using the *Survey Operations Manuals* (TIMSS, 1994d, 1994e) and *Sampling Manual* (TIMSS, 1994a) provided by the International Study Center. Only 7 of the 43 NRCs interviewed reported selecting a sample for any of the populations being tested without reference to the *Survey Operations Manuals* and the *Sampling Manual*. Explanations for deviations from the standard procedure tended to be practical in nature: for example, all schools were included in the population for that sample; national circumstances necessitated a change in sampling procedure; or someone other than the NRC was responsible for sampling.

About a third of the NRCs interviewed indicated that they had used the sampling and operations software provided by the International Study Center in order to facilitate the selection of classes and students. In some of these cases, NRCs found it convenient to use the software for one population but not another. Most of the NRCs (25 out of 43) reported using either their own software or other software such as Microsoft Excel or SAS programs that had been developed during the field trial. One NRC reported following the steps as outlined in the software but doing the actual sampling on paper instead.

In terms of the complexity of the procedures and number of personnel needed, most of the NRCs found the process of sample selection to be "somewhat difficult" or "not difficult at all." In the few cases where NRCs indicated the process was "very difficult" it was mainly because of a lack of resources, i.e. materials, staff, and funding.

Table 3.1
Interviews with National Research Coordinators

Country	Interview with NRC	Country	Interview with NRC
Argentina	X	Japan	X
Australia	X	Korea[2]	-
Austria	X	Kuwait[2]	-
Belgium (Fl)	X	Latvia	X
Belgium (Fr)	X	Lithuania	X
Bulgaria	X	Mexico	X
Canada (Alberta)	X	The Netherlands	X
Canada (Ontario)	X	New Zealand	X
Columbia	X	Norway	X
Cyprus	X	Philippines	X
Czech Republic	X	Portugal	X
Denmark	X	Romania	X
England	X	Russian Federation	X
France	X	Scotland	X
Germany[1]	-	Singapore[2]	-
Greece	X	Slovakia	X
Hong Kong	X	Slovenia	X
Hungary	X	South Africa	X
Iceland	X	Spain	X
Indonesia	X	Sweden	X
Iran	X	Switzerland	X
Ireland	X	Thailand	X
Israel	X	United States	X
Italy	X		

X = Interview Conducted
Total = 43

[1] Germany was unable to nominate a quality control monitor.

[2] Because of the timing of the funding of the quality assurance program, interviews with NRCs were not conducted in Korea, Kuwait, and Singapore.

3.4.2 WORKING WITH SCHOOL COORDINATORS

As the role of School Coordinator was vital to the successful implementation of the study, one function of the interview with the NRC was to assess the "readiness" of the School Coordinators in these countries.

At the time the interviews with NRCs were conducted, the majority of NRCs (38 out of 43) indicated that all the School Coordinators for their samples had been contacted, and that most NRCs (31) had already sent materials about the testing procedures to them. About half of the NRCs interviewed further indicated that they had already had formal training sessions for the School Coordinators.

3.4.3 TRANSLATING THE DOCUMENTS

The translation process and its verification was a major task for most participants. The interviews with NRCs attempted to assess whether any major problems were encountered that had not been previously exposed or dealt with.

Slightly over half of the NRCs found the process of translating and/or adapting the test booklets to be "somewhat difficult," compared with a third of the NRCs (14) reporting that the process was "not difficult at all." Of these 14 NRCs, 8 did not need to translate the documents since the international versions were prepared in English. When asked whether they used their own staff or outside experts to translate the booklets, most (24) reported that they used a combination of the two.

Thirty-three NRCs went through the recommended procedure of submitting their test booklets and receiving a Translation Verification Report from the International Study Center. Eight NRCs reported that they had not gone through this process, mainly because of time constraints. At the time of the interviews, one NRC had yet to receive the Translation Verification Report from the internationally commissioned reviewer.

NRCs generally found the process of adapting the questionnaires to be "somewhat difficult." Nine of the 10 NRCs that described this process as "very difficult" commented that many of the questions on the questionnaires were inappropriate for their country's educational system.

Adapting the *Test Administrator Manual* (TIMSS, 1994f), on the other hand, appears to have been a much easier process. Twenty-six of the NRCs indicated that the process was "not difficult at all." Eleven found the process "somewhat difficult", and only three considered the process to be "very difficult." Results were similar for adapting the *School Coordinator Manuals* (TIMSS, 1994b, 1994c). Most (24) had no trouble. Some (8) found the process to be a little difficult, and only a few (4) found it to be particularly difficult. Less than half of the NRCs interviewed (19) by quality control monitors either had translated or planned to translate the *Coding Guides for Free Response Items* (TIMSS, 1995a, 1995b) at the time of the interview.

3.4.4 ASSEMBLING AND PRINTING THE TEST MATERIALS

The procedure for the assembly of the test books was specified in detail in the *Survey Operations Manuals* (TIMSS, 1994d, 1994e) and in *Instructions for Preparation of the Instruments at the National Center.*

The assembly of the test booklets appears to have gone well throughout the study. Only two NRCs reported not being able to assemble the booklets according to the instructions provided by the International Study Center. One of these preferred to divide the test items into two books, and the other changed the number system for Population 1 in order to avoid potential confusion. Most of the 43 NRCs interviewed (30) experienced no difficulties actually assembling the test booklets. Only two NRCs indicated that the process was very difficult. Comments revealed that much of the difficulty was due to shortages of personnel and time.

Thirty-three of the NRCs interviewed reported conducting quality assurance procedures for checking the test booklets during the printing process. Three commented that this would be done by the printers; one pointed out that the process was not yet complete in that country; one indicated that the check would be performed before packing the materials; and two alluded to problems of staff shortage and lack of time. Several of the NRCs that did in fact conduct quality checks during printing discovered errors. The most frequently reported concern was "printing quality" followed by "pages missing" and "page ordering."

Most of the printing of test booklets and questionnaires was done by outside printers as opposed to in-house staff. Even so, only four of the NRCs interviewed reported not having followed specific procedures to protect the security of the tests during the assembly and printing process. Generally, the reasons given indicate that these NRCs considered such measures either unnecessary or not practical given their situation. Only one NRC found that the potential for a breach of security existed, but no details were provided.

3.4.5 PACKING, SHIPPING AND RETURNING THE TESTING MATERIALS

The *Survey Operations Manuals* (TIMSS, 1994d, 1994e) provided detailed instructions to the NRC for distributing and collecting the testing materials. There were specific instructions about what should be in each school's package and how the packages were to be assembled.

Very few errors were detected in packaging the materials for shipment to schools. Only 15 NRCs indicated that any errors were detected, and these tended to be minor and easily corrected. After distribution of materials, only 7 NRCs reported finding errors. At the time of the interviews, about half (21) of the NRCs indicated that they either planned to establish or already had established a procedure requiring schools to confirm receipt of the testing materials and verification of the contents. Concerns about confidentiality prevented about half of the NRCs from putting student names on the booklet covers.

3.4.6 CODING FREE-RESPONSE QUESTIONS

The selection and training of coders for the free-response questions was yet another vital component of the study and major task for the NRCs.

When asked who would primarily be coding the student responses to the free-response questions, most NRCs replied that this would be done by a combination of their own staff, teachers, and university students. The number of coders NRCs planned to use ranged between 4 and 65, with most NRCs using 20 or fewer. Three-quarters of the NRCs reported that at the time of the interview they had already selected the coders for the free-response items. Of these, many (19) had already trained the coders and scheduled the coding sessions for the free-response questions. Virtually all of the NRCs reported that they understood the procedure for coding the 10% reliability sample as explained in the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995c).

3.4.7 DATA ENTRY AND TRANSMITTAL

About half of the NRCs interviewed (21) indicated that they planned to use a combination of their own staff and outside experts to enter the data from the achievement test booklets and questionnaires into computer files. Most of those who had selected their data entry staff at the time of the interview (22) had already conducted training sessions. Twenty-eight of the 43 NRCs interviewed further planned to enter a percentage of test booklets twice as a verification procedure. That percentage ranged from 1% to 100%, with most of the NRCs reporting that they would double-enter between 6% and 10% of the data. Thirty-nine of the 43 reported that they had established a secure storage area to keep the tests following the coding of the responses.

3.4.8 QUALITY ASSURANCE SAMPLE

The NRCs were also responsible for conducting quality assurance observation visits in a tenth of the schools sampled. At the time of the interviews, approximately half of the NRCs had already selected their quality assurance sample for their on-site classroom observations. In most of the cases, the persons selected to do the observations were members of the NRC's staff. Several NRCs also relied, in whole or in part, on external agencies to complete this task.

3.5 SELECTION OF SCHOOLS FOR CLASSROOM OBSERVATION

Following the interview with the NRC, the quality control monitor and the NRC worked together to select 10 schools for classroom observation, plus 3 extra schools as potential replacements. Using the School Tracking Form, the quality control monitor and NRC chose the schools by a random selection process (albeit one subject to a number of practical constraints). The schools selected for classroom observation had to be within easy traveling distance of the quality control monitor's home so that travel and observation could be done in one working day; the NRC or quality control monitor had to be able to contact the school to ascertain the date and time of testing and to arrange the visit; the school could not be taking part in the NRC's own national quality control observation program; and the testing could not yet have taken place in that school. After the schools, the classrooms for observation were selected. Where possible, the class chosen was the upper-grade class. The school name and classroom to be observed were recorded on the Classroom Observation Tracking Form.

3.6 NATIONAL VERSIONS OF DATA COLLECTION INSTRUMENTS

At the end of the visit to the national center, the quality control monitor collected the following materials from the NRC:

- Test Administrator Manual (TIMSS, 1994f)
- School Coordinator Manual (TIMSS, 1994b, 1994c)
- Test booklets (for each population assessed)
- Performance assessment tasks (for each population assessed, if participating)
- School questionnaires (for each population assessed)
- Student questionnaires (for each population assessed)
- Teacher questionnaires (for each population assessed)
- Translation Verification Report (if this was not given to the quality control monitor at the training session)
- Student Tracking Forms for each class selected for observation
- Class Tracking Forms for each school selected for observation.

Quality control monitors received the Translation Verification Report either from the International Study Center during training or from the NRC on their visit to the national center. The quality control monitor checked that any deviations in translation or booklet layout were corrected before test administration, recorded that information, and submitted it to the International Study Center together with the instruments and manuals collected from the NRC.

3.7 SURVEY ACTIVITIES REPORT

NRCs were required to prepare a report on the survey activities and to submit the report to the IEA Data Processing Center together with their data files and documentation. The following indicates some important points this report was to cover.

- A description of the procedure used and any problems encountered in the translation, layout, and printing of the test instruments
- A description of any modifications in the international coding schemes
- An indication of which of the within-school sampling procedures applied to each population
- The national definition of mathematics classes, mathematics and science teachers, and streams (or tracks) that was used for the within-school sampling
- The criteria and definitions that were used for excluding students from testing within the selected schools

- If countries tested in languages other than English, the documentation of the coding schemes in the student data file
- A description of the problems encountered in the use of the survey tracking forms
- An indication of the procedures used for obtaining cooperation from schools and an indication of problems encountered
- Information on the organization of the testing sessions
- An indication of the position of school coordinators and test administrators in the schools
- A summary of the problems reported by the test administrators in the Test Administration Forms
- A description of any discrepancies in the timing of the testing sessions between the Test Administration Forms and the international instructions
- An indication of the procedures that were used for quality control and a summary of the findings from the national quality control monitors
- A description of the arrangements used for coding the student responses to the free-response questions, and problems encountered
- A description of data entry and data verification problems encountered, including an indication of error rate found during the verification of double-entered data
- Anything else that might help in interpreting the data or explaining possible anomalies.

NRCs were also required to submit a set of national survey instruments and a report on the appropriateness of the test items used in TIMSS. Forty-three reports were submitted. The information provided is summarized below.

Many of the NRCs reported following the TIMSS guidelines on translation, layout, and printing. Countries reported following TIMSS guidelines when translating from English to the country's local language. For English-speaking countries, only minor adaptations in the items were made. These changes generally reflected regional differences in spelling (e.g. colour for color), and regional name preference. Translation of the TIMSS documents was more problematic. Several NRCs commented upon the difficulty of making exact translations of single words or phrases from English into the country's language. In addition, NRCs reported attempts to ensure that items were presented in a manner that gave a "natural level for the grade in question." While a number of countries stated that they were able to have their translated instruments verified before printing the test booklets, some NRCs reported that due to a lack of time they could not have their translations verified prior to printing. Other NRCs did not mention the international verification process.

Frequent problems reported with respect to the test booklets were missing pages, blank pages, and duplicated pages, in addition to problems associated with item translation. Several NRCs commented that the timeline did not allow for a more thorough review of the test booklets before printing and dissemination. In general, where it was necessary to translate the questionnaires and manuals into another language, the NRCs felt that more time was needed. In a few cases, mention was made of the cost of printing and the need to obtain financial support and help in getting the booklets duplicated.

All NRCs reported making adaptations in the international test booklets. In some countries, items were adapted by a committee, while in others, the task was completed by one expert. Most countries reported making only minor, mainly name, changes in the test items. Only one country mentioned a severe mismatch between test booklet items and curriculum. Another country mentioned that several items referred to human reproduction and noted that its request to TIMSS that these items be removed had not been heeded.

With few exceptions, NRCs reported using within-school sampling Procedure A, based on selecting intact classes. In some cases, a random sample of students was selected from all the students in a particular year level and intact classes were not used. Many NRCs reported that no streams or tracks were used in their selected classes. In some cases, students were tracked by the school system within which they were placed. Few countries reported that streams or tracks were important to their school system.

Excluded students either had some disability or parental permission had not been given for their participation in the study. For example, several countries reported excluding functionally disabled students, educable mentally retarded students, students not speaking the native language, and students having subject- or reading-specific difficulties. One country reported that some students were excluded from the study because parental permission for their participation in the study had not been given.

The time allocated to the achievement tests was found to be too generous by many NRCs (64 minutes for Population 1, 90 minutes for Populations 2 and 3). Many students had completed their test in 30 minutes. In contrast, students needed 10 to 20 minutes more than expected for filling out the entire questionnaire, and many students were "very tired at the end of the session and were not able to concentrate fully on their work."

Some NRCs mentioned that the study was affected by events outside their control. For example, in a few cases strikes affected communication between NRCs and schools. In some cases schools were reluctant to participate in the study since they were trying to meet local demands or were already participating in other studies. Comment was also made on the heavy demands made by the TIMSS study. For example, the teacher questionnaire was criticized as being "too elaborate, overloaded, too long, too detailed, and consuming too much time."

The procedures for obtaining cooperation from schools, and the problems encountered, fall into two main categories. In some countries, participation was under the control of a central authority, such as a ministry of education, and participation levels and cooperation levels were very high. In other school systems, participation in the study was voluntary, with the schools themselves deciding whether or not to participate. One NRC reported that, despite letters from the director of education, the help of the education department, and the staff of the faculty of education of a university, the biggest problem of the study was getting schools to participate. One NRC obtained a very high participation rate by asking ministerial inspectors of mathematics to function as quality control managers. Several NRCs commented favorably upon the support they received from their school systems. For example, one NRC commented that "we would like to express our appreciation for the good cooperation between selected schools, County Inspectorates, and the National Center."

3.8 SUMMARY

In order to carry out the International Study Center's quality assurance program, quality control monitors were hired and trained for each participating country. Each quality control monitor was provided with a procedural manual and data-collection instruments, and was required to visit the TIMSS national center to (1) interview the NRC about aspects of their data collection activities, (2) work with the NRC to select a sample of schools to visit, and (3) collect copies of national versions of the TIMSS data collection instruments.

The results of the interviews indicate that NRCs had generally prepared well for the data collection, and, despite the heavy demands of the schedule and shortages of resources in some centers, were in a position to conduct the data collection in an efficient and professional manner. Quality control monitors succeeded in selecting schools for their visit to observe a test administration session, and collected copies of the TIMSS instruments as requested.

Following the completion of their data collection, NRCs were asked to submit a report describing their experiences and documenting any unusual occurrences or deviations from prescribed procedures. Most NRCs complied with this request, and while many minor mishaps and delays were recorded, in general most NRCs managed to follow the TIMSS procedures and collect their data in a satisfactory manner.

REFERENCES

Third International Mathematics and Science Study (TIMSS). (1994a). *Sampling Manual–Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *School Coordinator Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *School Coordinator Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *Survey Operations Manual– Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994f). *Test Administrator Manual–Populations 1 and 2.* (Doc. Ref.: ICC890/NRC426). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897/NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b*). Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC913/NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995d). *Manual for the TIMSS Quality Control Monitors* (Doc. Ref.: ICC920/NRC450). Chestnut Hill, MA: Boston College.

4. OBSERVING THE TIMSS TEST ADMINISTRATION

Michael O. Martin
Craig D. Hoyle
Kelvin D. Gregory

4.1 OVERVIEW

In order to monitor compliance with international procedures in the administration of the TIMSS tests, quality control monitors visited a sample of schools where they observed testing sessions and interviewed School Coordinators. Table 4.1 summarizes the number of test administrations observed in each country. The complete program (visit to a school, observation of the test administration, and an interview with the School Coordinator) was implemented in 37 countries. The program was implemented only partially or not at all in the remaining countries for a variety of reasons.

- Because of the timing of the funding of the quality assurance program, countries on the Southern Hemisphere timeline for Population 1 and 2 (Australia, New Zealand, Republic of Korea, and Singapore) had already completed their testing program and therefore test sessions could not be observed by the quality control monitor. Australia and New Zealand carried out the program with samples of Population 3 testing sessions.
- Several other countries were scheduled to have completed testing before a quality control monitor could visit for classroom observations (Iceland, Japan, Kuwait, and Thailand). Iceland did conduct classroom observations for Population 3, but the materials arrived at the International Study Center too late to be included in this report. Japan and Thailand conducted interviews with samples of School Coordinators after testing had taken place.
- Denmark and Ireland, both countries where school participation was voluntary and testing was conducted by classroom teachers, were unwilling to ask schools to take part in the program of classroom testing observations.
- In the United States, the quality control monitor became indisposed, and was unable to conduct the classroom observations. Information about the testing sessions in the selected schools was collected at a later date from Test Administrators and School Coordinators.
- Germany was unable to nominate a quality control monitor in time to observe testing sessions.

Table 4.1
Classroom Observation Records

Country	Number of Observation Records
Argentina	10
Australia	8
Austria	10
Belgium (Fl)	10
Belgium (Fr)	10
Bulgaria	10
Canada (Alberta)	10
Canada (Ontario)	10
Columbia	10
Cyprus	10
Czech Republic	10
Denmark[1]	-
England	10
France	11
Germany[2]	-
Greece	10
Hong Kong	7
Hungary	10
Iceland[3]	-
Indonesia	11
Iran	10
Ireland[1]	-
Israel	10
Italy	10
Japan[4]	6
Korea[3]	-
Kuwait[3]	-
Latvia	7
Lithuania	10
Mexico	10
The Netherlands	10
New Zealand	9
Norway	10
Philippines	10
Portugal	10
Romania	10
Russian Federation	10
Scotland	10
Singapore[3]	
Slovakia	10
Slovenia	10
South Africa	9
Spain	9
Sweden	10
Switzerland	10
Thailand[4]	5
United States[5]	12
Total = 384	

[1] Unwilling to take part in classroom observations.

[2] Unable to nominate a quality control monitor to observe testing.

[3] Tests conducted before quality control monitoring programs were in place.

[4] Unable to conduct observations but did conduct interviews with School Coordinators.

[5] Unable to conduct observations but did conduct interviews with Test Administrators.

During each visit to a school, the quality control monitor documented the information he or she collected in a questionnaire called the Classroom Observation Record. This had four sections:

- Activities preliminary to the testing session, including preparation, test security, arranging accommodation

- Observation of the testing session
- Quality control monitor's general impressions of test administration
- Interview with the School Coordinator.

This chapter provides a general summary of the results of the school visits as reported by the quality control monitors. Detailed results are presented in Appendix E. The letter/number codes displayed after the headings in the commentary below correspond to the instrument questions and results provided in the appendix.

4.2 SCHOOL VISITS AND TEST SESSION OBSERVATIONS

4.2.1 SECTION A : PRELIMINARY ACTIVITIES OF THE TEST ADMINISTRATOR

Having become acquainted with the Test Administrator and located the room where the testing session would take place, the quality control monitor was to record observations on the condition of the testing materials, the Test Administrator's level of preparation, and the suitability of the testing room.

Overall, the quality control monitors reported very favorably upon the preliminary activities conducted by the Test Administrators. With very few exceptions, test conditions, booklets, and directions were in accordance with the study procedures. Where changes were made, they tended to be minor in nature and posed no threat the validity of the study.

- *Verification of the supply of test books (A.1).* Almost all Test Administrators (94%) had definitely (77%) or probably (17%) verified adequate supplies of test books prior to the test administration.
- *Seals on test books (A.2).* In every session where the national center had used booklet seals to enhance booklet security (128 of the 384 sessions observed), seals were intact before the testing session began.
- *Test booklet and Student Tracking Form correspondence (A.4).* Student identification on the test booklets and questionnaires corresponded to the information on the Student Tracking Form in almost every session (96%). In the remaining sessions, there were either minor problems involving only a few students or changes in procedure such as filling out the Student Tracking Form after the tests were distributed.
- *Correct version of the administration script (A.6).* In practically all sessions (99%), the Test Administrator had the correct version of the administration script for the session.
- *Familiarization with script (A.7).* Most Test Administrators (93%) had definitely (74%) or probably (19%) familiarized themselves with the administration script before the testing session.
- *Space for students to work (A.8).* In most sessions (93%), there was adequate seating space for the students to work without distractions. Comments from quality control monitors indicated that seating arrangements varied considerably, but that, for the most part, seating arrangement was not a problem. In a few sessions, however, it is clear that seating arrangements were less than ideal.

- *Adequate room for the Test Administrator to circulate (A.10).* In most of the sessions (97%), the Test Administrator had adequate room to move about during the testing session and ensure that students were following directions correctly.
- *Keeping track of time (A.12, A.14).* Most Test Administrators (90%) had a stop-watch or timer for accurately timing the testing sessions. Some of those who did not may, however, have had an ordinary watch. The presence of a wall clock for students to keep track of time was more the exception than the rule. Quality control monitors reported the presence of wall clocks in only 26% of the sessions. No data were collected, however, on the number of students with their own watches.
- *Supply of pencils (A.13).* The Test Administrator had an adequate supply of pencils and other necessary materials ready for the students in 77% of sessions. However, in many countries it is the responsibility of the student to bring pencils, pens, etc., to testing situations, so it is not necessarily the case that there was an inadequate supply in almost a quarter of the testing sessions.

4.2.2 SECTION B: OBSERVING THE TESTING SESSION

Following the preliminary activities, the quality control monitor was required to observe the testing session, and record the activities of the Test Administrator throughout the session. In Population 1 and 2 schools, a test administration consisted of two testing sessions, Session 1 and Session 2, separated by a short break, followed by a session for the student questionnaire after a second break. In Population 3 schools, there was a single testing session, followed after a break by a session for the student questionnaire.

In many cases, quality control monitors reported no changes in the script. Where changes were observed, they tended to be local adaptations. The testing sessions were orderly and well conducted. The time allowed for testing was generous, and few quality control monitors reported tests extending past the scripted time. Approximately half of the quality control monitors reported that more time was required to complete the student questionnaire.

SESSION 1 (ALL POPULATIONS)

- *Following the Administrator's Script (B.2a, B.2b, B.2c).* Most Test Administrators (95%) followed the instructions for preparing students in the administrator's script without making any changes (64%) or making only minor changes (31%). Most (97%) also followed the script with regard to distributing the materials, making no changes (83%) or minor changes (14%). Likewise, a very high percentage (97%) had made either no changes (82%) or only minor changes (15%) in the instructions to begin testing.
- *Changes to the script that were made (B.3a).* Approximately half of the Classroom Observation Records indicated that no changes were made in the script. Where changes were made, the Test Administrator essentially adapted the script in a manner pertinent to the students. No quality control monitor reported deviations that might be expected to affect the interpretation of test results.
- *Distribution of the test books (B.4).* In almost every session (97%), test booklets were distributed one at a time, as prescribed in the manual. In those few sessions where the procedure was not followed, some other acceptable method (in the judgment of the quality control monitors) of distribution was used.

- *Allocation of test books (B.6, B.7).* In most sessions (93%), test booklets were distributed according to the booklet assignments on the Student Tracking Form. In some of the sessions where they were not, it was because the Student Tracking Form was not available at the time of the session. Other cases involved one or two students who were issued spare booklets or whose booklet ID did not match the Student Tracking Form. In each of these cases, the actions of the Test Administrator were recorded on the Student Tracking Form.
- *Attendance (B.8).* The Test Administrator recorded attendance correctly on the Student Tracking Form in almost all sessions (98%).
- *Testing time (B.9, B.10., B.11).* In most test administration sessions (86%), the appropriate amount of testing time was allocated to Session 1. In many sessions, however, all of the students finished the test before the prescribed time had elapsed, and so the test administrator brought the session to an early conclusion.
- *Time announcements (B.12, B.13).* In most sessions (86%), the Test Administrator announced "you have 10 minutes left" prior to the end of Session 1, as instructed. Quality control monitors indicated that in 24% of the sessions other announcements regarding the time remaining were made during Session 1.
- *Instructions to stop work (B.15).* In almost every session (98%), students complied very well (77%) or well (22%) with the instructions to stop work.
- *Collection of the booklet at end of Session 1 (B.16).* Many Test Administrators chose not to collect the test booklets at the end of Session 1. Only 51% followed the prescribed procedure of collecting test booklets one-at-a-time from each student. In many cases, test booklets were left on the students' desks between sessions. Sometimes this was necessary as the test booklets and questionnaires were in a single packet. None of the quality control monitors recorded any observations that would indicate that test integrity had been compromised.

SESSIONS 2 AND 3 (POPULATIONS 1 AND 2)

- *Break between Session 1 and Session 2 (B.19, B.20, B.21).* The recommended break time between testing sessions was 20 minutes, although in the majority of sessions (81%) some other interval was found to be more convenient. Frequently, breaks coincided with lunch or recess periods. In some instances, there was no break between sessions; in others, the break time was substantially shortened. In most sessions (84%), however, despite changes in its length, the break was conducted exactly (56%) or almost (28%) as prescribed.
- *Session 2 restart time (B.22, B.23, B.24).* Most sessions (54%) required less time than the prescribed five minutes to re-read instructions and settle students at the beginning of Session 2. Explanations for the deviation from the scripted 5 minutes included "the students had no questions," and "students embarked immediately on the second part of the test."
- *Testing time session 2 (B.25, B.26, B.27).* For most sessions (79%), the testing time for Session 2 was as prescribed in the administrator's script. As with Session 1, Test Administrators sometimes brought the session to an early close if all students finished before the prescribed time had elapsed.
- *Time remaining announcement (B.28, B.29, B.30).* Generally the Test Administrators (81%) announced "you have 10 minutes left" prior to the end of Session 2, as prescribed in the manual. In most instances where this announcement was not made, an acceptable procedure was substituted. Such a substitution was made in about 20% of the sessions.

In several sessions, administrators kept track of time by marking off intervals on a blackboard.

- *Ending Session 2 (B.31).* In almost all sessions (97%), students complied with the instruction to stop work either very well (81%) or well (16%).
- *Collection of test booklets after Session 2 (B.32).* In most cases (81%), the Test Administrator collected the test books one at a time at the end of Session 2. Where the books were not collected as scripted, the test administrator used an alternative method that did not compromise the integrity of the test administration.
- *Announce break before student questionnaire (B.34).* In two-thirds of the sessions, the Test Administrators announced a break at the end of the testing session, to be followed by the student questionnaire. In many of the remaining sessions, the administration of the student questionnaire followed without a break. Sometimes the student questionnaire did not follow the testing sessions immediately but was completed on another occasion.
- *Read script for break (B.36, B.37).* In most of the sessions (89%), the Test Administrators followed the script to end the testing and signaled a break either verbatim (63%) or with minor changes (26%). Minor changes in the script were noted in 44 sessions, with additions in 24 sessions and omissions in 38 sessions (some sessions had both additions and omissions).
- *Break conducted (B.38, B.39, B40, B.41).* Most (82%) of the Test Administrators held the break as directed in the manual (68% exactly as directed; 14% nearly as directed).
- *Distribution of the student questionnaire (B.43).* The majority (67%) of Test Administrators distributed the student questionnaire and gave directions as specified in the script. In many countries, the student questionnaire was distributed in a packet with the test booklets, not separately. There was no indication of any problems with the distribution of the student questionnaires.
- *Time allocated to Student Questionnaires (B.46, B.47, B.48, B.49).* In more than half the sessions (60%), the student questionnaire required more time than was prescribed in the administration script. The Test Administrator Manual made provision for more time for the questionnaire as necessary. Extra time allowed ranged from 1 to 45 minutes, with a median of 20 minutes.
- *End of session (B.50, B.51).* In 80% of the observed sessions, the Test Administrator thanked students for participating in the study. Dismissal of students was generally an orderly affair. Quality control monitors described 94% of the session dismissals as either very orderly (62%) or somewhat orderly (32%).

SESSION 2 (POPULATION 3)

- *Break announcement after testing session (B.69).* Test Administrators announced a break at the end of 64% of the testing sessions, to be followed by the student questionnaire. In many of the remaining sessions the administration of the student questionnaire followed without a break. Occasionally the student questionnaire was completed at another time.
- *Script (B.71, B.72).* For most sessions (91%), Test Administrators followed the script to end the testing and signal a break either verbatim (52%) or with minor changes (39%). Minor changes in the script were noted in 12 sessions, with additions in 2 sessions and omissions in 9 sessions (some sessions had both additions and omissions).
- *Break time (B.73, B.74, B.75, B.76, B.77).* The break time differed from the time recommended in the script in 42% of the questionnaire sessions. Most of these (33% of all sessions) involved a shorter break time. Most (77%) of the Test Administrators

conducted the break as directed (66% exactly as directed; 11% nearly as directed). The most common reason given for not including a break was that the country's Test Administrator Manual did not provide for one.

- *Distribution of Student Questionnaires (B.78, B.79).* Test Administrators distributed the Population 3 student questionnaire and gave directions as specified in the script in 71% of sessions. As in Populations 1 and 2, in many countries the student questionnaire was distributed in a packet with the test booklets. There was no indication of any problems with the distribution of the student questionnaires.
- *Time allocated to Population 3 student questionnaires (B.80, B.81, B.82, B.83, B.84).* There was considerable variation in the amount of time required to complete the Population 3 student questionnaire. In 40% of sessions, quality control monitors reported that the time allowed was less than the prescribed amount, whereas in 23% of the sessions, additional time was requested.
- *Dismissal of Population 3 students (B.85, B.86).* Most (82%) of the Test Administrators thanked students for participating at the end of the study. Dismissal of students was described as "very orderly" (75%) or "somewhat orderly" (20%).

4.2.3 SECTION C: SUMMARY OBSERVATIONS OF THE QUALITY CONTROL MONITORS

Following observation of the testing session, quality control monitors were asked to give their impressions of several aspects of the test administration, including the behavior of the students and the activities of the Test Administrator.

With few exceptions, quality control monitors commented favorably on test administrations. They stated that the Test Administrator conducted the test sessions in a well organized and professional manner. They found that students were well motivated and challenged by the test items. Where the quality control monitors noted deviations from the administration script, these deviations posed little if any threat to the validity of the results. Rather, the changes mostly represented acceptable adaptations in the test administration.

- *Student conduct (C.1, C.2).* In 94% of the sessions, students were described as either extremely (65%) or moderately (29%) orderly and cooperative. In the rare situations where students were not cooperative, quality control monitors indicated that the Test Administrator almost always made some effort to exert control.
- *Supervision by the Test Administrator (C.3).* Test Administrators walked around the room to monitor student behavior in 94% of the observed sessions. Where this did not occur, it was often because of lack of space.
- *Student Questions (C.5).* Test Administrators addressed students' questions appropriately in almost all sessions (97%).
- *Evidence of cheating (C.7).* In most sessions (87%), there was no evidence of students attempting to cheat. Where evidence was reported, it was usually that students attempted to talk to their neighbors or attempted to copy from a neighbor's booklet. Because eight different booklet versions were in use in a classroom, copying responses from a neighbor's booklet was unlikely to help a student's performance.

- *Defective booklets (C.9, C.10, C.11).* In 6% of the observed sessions, defective test booklets were identified and replaced before the session began. In a further 6% of sessions, defective booklets were found and replaced after the sessions began. On most of the occasions where booklets needed replacement, the Test Administrator replaced them appropriately. Occasionally booklets were not replaced because of a lack of spare copies.
- *Late students (C.13).* In most sessions (88%), no late students were reported. In 2% of sessions, late students were not admitted; in 5% of sessions, late students were admitted before the testing began; and in 5% of sessions, they were admitted after testing had begun.
- *Refusals to take the test (C.14, C.15, C.16).* In just 3% of sessions did students refuse to take the test, and then usually just one student. In only one session was it reported that more than a few students refused to take the test. This case is described in more detail below. Test Administrators accurately followed the instructions for excusing students in 5 of the 9 sessions where this was necessary. In the single case where more than five students were excused, the entire class refused to take the test. The quality control monitor noted that the students were all of low ability and wanted to give up because the test was too difficult for them, and that the Test Administrator had persuaded them to attempt the test. In addition, the monitor noted that the school was for students who could not gain entry to "good schools."
- *Emergency during testing (C.17, C.18).* In 15% of sessions, at least one student left the room during testing because of an "emergency." Usually the "emergency" was merely a need to visit the bathroom.
- *Overall quality of the test administration session (C.19, C.20).* The overall quality of the testing sessions was rated high, with 94% of sessions rated "good" or better. The Test Administrator usually was praised, as were the students. It was commonly observed that students were well disciplined, motivated, and challenged by the test. In many cases, quality control monitors noted that the Test Administrator had conducted the test in a well organized, professional manner. Critical comments by monitors focused predominantly upon the time allocated to the test and the language used in the test, and not upon the actual test administration.

4.2.4 SECTION D: INTERVIEW WITH THE SCHOOL COORDINATOR

Following the completion of the testing in the school, quality control monitors were asked to interview the School Coordinator to collect information on experiences with the test administration, attitudes and reactions of school staff, and suggestions for improvements for the future.

The comments of the School Coordinators tended to be very positive. They were happy with the shipping of TIMSS materials, and overwhelmingly made positive comments regarding the National Research Coordinators. Negative criticisms centered mainly upon the teacher questionnaire, the mismatch between test items and curriculum, and the timing of the testing program.

- *Overall impression (D.1).* School Coordinators almost unanimously (99%) indicated that the testing sessions went well (70% very well).

- *Attitude of other school staff (D.2).* Most School Coordinators (71%) rated the attitude of other school staff members towards the TIMSS testing as positive. Negative attitudes (4%) were predominantly attributed to the date of testing, which caused disruptions in the regular schedule. A further 25% of the School Coordinators rated the attitude of other school staff as neutral to the TIMSS testing.
- *Checking materials (D.3).* Most School Coordinators (87%) found time to check the shipment of materials from the National Research Coordinator prior to the day of testing.
- *Items received (D.5).* In most cases, School Coordinators reported receiving the correct shipment of test booklets (99%), Test Administrator Manuals (100%), School Coordinator Manual (98%), Student Tracking Forms (93%), Student Questionnaires (98%), Teacher Questionnaires (91%), School Questionnaires (99%), and Test Administration Forms (92%). Teacher Tracking Forms (70%), Student-Teacher Linkage Forms (26%), and envelopes or boxes for the purpose of returning the materials after the assessment (71%) were less frequently reported to be correct, but in some instances these items may have been purposely omitted by the NRC.
- *Responsiveness of National Research Coordinators (D.6).* School Coordinators felt that the National Research Coordinator was responsive to questions and concerns in most (93%) cases.
- *Collecting teacher questionnaires (D.7, D.8).* In many schools (60%), it was not possible for the School Coordinator to collect completed teacher questionnaires before the test administration. These usually were completed during or after the test administration, with several observations noting that Test Administrators were unaware that the questionnaire was to be collected before the test administration. Many of the teachers (60%) commented that the questionnaire took more time than expected to complete.
- *Satisfaction with testing room (D.11).* Most School Coordinators (96%) were satisfied with the testing room that they were able to arrange for the testing session.
- *Make-up-sessions (D.12, D.13).* Most School Coordinators (84%) anticipated that make-up sessions would not be required at their school. Most (93%) of those who anticipated the need for make-up sessions planned to conduct one.
- *Selection and training of Test Administrators (D.14).* School Coordinators predominantly made positive comments regarding the training of Test Administrators. The *Test Administrator Manual* was generally found to be very useful. In some cases, they suggested improvements such as adding flow diagrams to the manual, and a more extensive training period.
- *Motivational talk (D.16).* Almost half of the School Coordinators (46%) reported that students received special instructions, motivational talks, or incentives to prepare them for the assessment. Most of these consisted of introductory presentations by the school principal, class teacher, or other test administrator.
- *Practice questions (D.18).* Only 2% of the School Coordinators reported that students were given an opportunity to practice on questions like those in the tests before the testing session.
- *School Coordinator Manual (D.20).* The majority of School Coordinators (92%) believed that the *School Coordinator Manual* worked well.
- *Completeness of class lists (D.23).* Most (93%) of the School Coordinators reported that the classes listed on the Class Tracking Form for the school represented a complete list of the mathematics classes in that school at those grades.

- *Students not in math classes (D.25).* School Coordinators almost universally (98%) reported that to the best of their knowledge there were no students in their schools at the required grade levels who were not in any of the mathematics classes listed on the Class Tracking Form.
- *Students in more than one math class (D.27).* Most School Coordinators (96%) also believed that there were no students in the required grade levels who were in more than one mathematics class.
- *Willingness to serve again (D.29).* Most of the School Coordinators (90%) indicated that if there were another international assessment, they would be willing to serve as School Coordinators again.

4.3 SUMMARY

In order to monitor compliance with international procedures in the administration of the TIMSS achievement tests, the International Study Center dispatched a quality control monitor to each country to visit a sample of schools where they observed a testing session and interviewed the School Coordinator. Test administrations were observed and School Coordinators interviewed in 37 countries, and interviews were conducted with School Coordinators or Test Administrators in three further countries.

The Classroom Observation Record completed by the quality control monitor for each school visit had four sections:

- Activities preliminary to the testing session
- Observation of the testing session
- Quality control monitor's general impressions of the test administration
- Interview with the school coordinator.

In general, quality control monitors reported very favorably on the test administration effort. Test Administrators were well prepared, and, with few exceptions, test conditions, instruments, and directions were in accordance with prescribed procedures. Test administrations were reported to be orderly and well conducted. The time allowed for testing was found to be generous, with very few reports of students needing more time. With very few exceptions, quality control monitors commented favorably on the test administrations. Generally, they reported that Test Administrators were well organized and performed their duties in a professional manner, and that students were orderly and applied themselves to their tasks. School Coordinators also tended to be very positive in their remarks. Despite the disruption to school schedules, school staff were generally reported to have favorable attitudes towards the project. The burden of completing the teacher questionnaire drew adverse comment from quite a few teachers.

On the evidence provided by the quality control monitors from their school visits the TIMSS test administration was generally a very successful endeavor. Readers and reviewers can be assured that the TIMSS data were collected following standard procedures and under standard conditions to the greatest extent possible.

5. QUALITY CONTROL STEPS FOR FREE-RESPONSE SCORING

Ina V.S. Mullis
Teresa A. Smith

5.1 OVERVIEW

For the TIMSS main surveys, approximately one-third of the written test time was devoted to free-response items, both short-answer and extended-response types. Across the seven tests administered to the three populations (mathematics and science at Populations 1 and 2, as well as literacy, advanced mathematics, and physics at Population 3) and the performance assessments administered to Populations 1 and 2, TIMSS included approximately 300 free-response questions and tasks.

The free-response items were scored using two-digit codes with rubrics specific to each item. The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code used to identify specific types of approaches, strategies, or common errors and misconceptions.

The scope of the free-response scoring effort was very complex. With large within-country samples of students responding to the tests, and those student samples representing many countries, ensuring reliability of scoring was a major concern for TIMSS. It was therefore necessary to develop procedures for applying the coding guides reliably and to document coding reliability.

To meet the goal of ensuring reliable scoring, TIMSS used a three-pronged approach.

1. An ambitious schedule of training sessions was designed to assist representatives of national centers who would then be responsible for training personnel in their respective countries to apply the two-digit codes reliably.
2. To gather and document information about the within-country agreement among coders, TIMSS developed a procedure whereby approximately 10% of the student responses were to be coded independently by two readers.
3. To provide information about the cross-country agreement among coders, TIMSS conducted a special study at Population 2 whereby 39 coders from 21 of the participating countries coded common sets of student responses.

This chapter contains information about these three activities. For more details about the training sessions and the procedures for estimating within-country reliability, please see:

- "Training Sessions for Free-Response Scoring and Administration of the Performance Assessment" (Mullis, Jones, and Garden, 1996)
- *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995).

5.2 TRAINING SESSIONS FOR FREE-RESPONSE SCORING

Training sessions for free-response scoring were conducted in seven regions to provide easier access for participants, and smaller groups for the TIMSS trainers to manage. Accommodations also were required to address the TIMSS schedule, which, for the most part, required countries on the Southern Hemisphere timeline to test Populations 1 and 2 in the fall of 1994 and Population 3 in the fall of 1995. The remaining countries tested all three populations in the spring of 1995. Consistency across sessions was provided by using essentially the same training team and training materials for all the sessions. The members of the training team had considerable knowledge of the TIMSS tests and of the procedures used in training scorers to achieve high reliability in scoring. The team consisted of the following members: Mr. Chancey Jones, United States; Mr. Robert Garden, New Zealand; Dr. Graham Orpwood, Canada; Dr. Jan Lokan, Australia; and Dr. Ina Mullis, United States. Although not all training team members attended all of the training sessions, most attended the larger sessions and at least some attended each of the sessions. Representatives from each of the participating countries attended at least one training session. The only exception was Italy, which to date has not had the resources to code and enter the data it collected. The schedule of training sessions and the countries participating in each session are shown in Table 5.1.

A four-day training schedule was developed to introduce attendees to the TIMSS coding approach and give them practice in scoring example papers. The specifics of the schedule varied from session to session depending on the participants' involvement in the different aspects of TIMSS. However, in the most effective schedule for the sessions, the first three days were devoted to scoring procedures for the main survey at Populations 1, 2, and 3, respectively, and the fourth day to the performance assessment. The sessions began with an orientation covering the importance of the coding of free-response questions and performance tasks. The training team described the significance of the first and second digits in the TIMSS codes, explaining that the first digit is a correctness score, and that the second digit provides diagnostic information about the type of response. Other orientation topics included the importance of maintaining high reliability in conducting the coding process, the desirability of planning and conducting similar training in the participants' own countries, and the necessity of finding exemplar student papers within each country to use in the training process. Information also was provided about procedures for conducting the actual coding and the within-country reliability studies (as described in the *Guide to Checking, Coding, and Entering the TIMSS Data*, TIMSS, 1995).

Table 5.1
TIMSS Free-Response Item Coding Training Sessions

Date	Location and Participating Countries
October 10-12, 1994	Wellington, New Zealand (Populations 1 and 2) — Australia, Korea, New Zealand, Singapore
January 18-21, 1995	Hong Kong — Hong Kong, Japan, The Philippines, Thailand
January 25-28, 1995	Boston, United States — United States, Canada, Mexico, Norway, Kuwait
March 7-10, 1995	Enschede, The Netherlands — Belgium (Flemish), Denmark, England, France, Germany, Greece, Indonesia, Iran, Ireland, The Netherlands, Portugal, Scotland, Spain, Sweden, Switzerland
March 13-16, 1995	Budapest, Hungary — Austria, Bulgaria, Canada, Cyprus, Czech Republic, Hungary, Iceland, Israel, Latvia, Lithuania, Norway, Romania, Russian Federation, Slovak Republic, Slovenia, the Ukraine
July 17-18, 1995	Miami, United States — Colombia, Argentina
July 18-19, 1995	Pretoria, South Africa — South Africa
September 6, 1995	Wellington, New Zealand (Population 3) — New Zealand
September 28-29, 1995	Melbourne, Australia (Population 3) — Australia

Each participant in the training sessions needed a considerable amount of material, including the relevant coding guides, manuals, and packets of example papers for practice. TIMSS developed an extensive coding guide for each population, containing the individual rubrics developed for each of the TIMSS free-response items given to that population. Each rubric defined the scoring categories to be used for the item together with example student responses for each category.

Training packets were prepared for a subset of the items considered the most complicated to score. Across the populations, training packets were prepared for 14 mathematics and 23 science items. Each packet began with the rubric for the item followed by coded student responses illustrating each of the categories in the rubric. The packet also contained about 15 to 20 precoded student responses, with the codes known to the training team but not to the participants in the training session. These were used to give the participants an idea of what it is like to actually score student responses.

The purpose was not to conduct the actual training for the coders, but to present a model for use in each country and present an opportunity to practice with the most difficult items. The trainers emphasized the need for participants to prepare training materials for each of the items rather than only a sample of items, and the fact that for more difficult items more example responses might be needed to help coders reach a high degree of reliability.

The trainer for the item would begin by familiarizing the group with the rubric for the item and answering questions about the reasons underlying the categories. Then the trainer would invite the participants to code five or six of the example student responses. After the group had completed the coding for these responses, the trainer would read the scores for the responses and answer any questions from the group. This procedure was iterated until all the precoded responses were scored by the participants. Although there was insufficient time at the training sessions to achieve a consistently high level of agreement on each of the items, the procedures provided some practice for participants and an example for how training might be conducted in each country.

Spending only one day on each of the three populations with a fourth day for countries participating in the performance assessment made for a demanding and intense session for most participants. In the future, it would be beneficial to devote more time to training in free-response scoring. All in all, however, the model of developing detailed coding guides and "training the trainers" appears to have worked successfully.

5.3 WITHIN-COUNTRY RELIABILITY STUDIES

In addition to using well-defined coding rubrics and careful training procedures, TIMSS also implemented procedures to monitor inter-rater reliability within each participating country. The procedures were designed to document the degree to which the same codes were given to the same responses regardless of the coder.

The TIMSS International Study Center recommended that each country use a method whereby 10% of the booklets would be coded independently by two coders. Explained in detail in the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995), the procedure called for every 10th booklet to be coded by two different individuals, with neither knowing the identity of the other or the codes assigned.

Because it is important that the booklets selected for the reliability study represent the coding process in general, the procedures for the reliability sample needed to be as routine as possible to blend in with the normal coding procedure. The object is for the reliability sample to provide an estimate of the overall quality of the free-response coding.

The general idea was to divide coders into two equivalent groups (Group A and Group B, balanced in terms of numbers, training, and experience) and to divide the booklets into two equivalent sets (Set A and Set B, according to odd versus even school identification numbers). The coders in Group A were to code all the booklets in Set A and the 10%

reliability sample of the booklets in Set B, while the coders in Group B coded all of the booklets in Set B and the 10% reliability sample of the booklets in Set A. Each group, therefore, handled both sets of booklets.

Because the coders could not know each other's codes for the reliability sample, ensuring a "blind" coding for the reliability sample necessitated the preparation of separate coding sheets for the 10% reliability sample. Coders were to handle the reliability set of booklets first, recording their results on a separate answer sheet. For the other set, the group coded all the booklets, and the codes were written directly into the booklets.

This procedure ensured that the coding of the reliability sample was conducted without the coders knowing the codes for the main survey and vice versa. It also ensured that different coders worked on the reliability sample than on the main coding, so that the same coder did not provide the codes for both reliability sample and main survey. As an additional step, countries were encouraged to try as much as possible to balance the reliability sample coding for each of the Group A coders across the different Group B coders, and similarly to balance the reliability sample coding for each of the Group B coders across different Group A coders. Countries also were encouraged to do the reliability scoring throughout the main survey coding. That is, for an hour or so each day, the Group B coders were to code every tenth booklet in the Set A batches, while the Group A coders were coding every tenth booklet in the Set B batches.

Many suggestions were given to countries about how to implement the 10% reliability scoring. Above all, however, the TIMSS International Study Center emphasized the importance of implementing a systematic plan to document the reliability of the coding schemes and stressed the need to enter the information about coding reliability into the database.

The within-country scoring reliability results for Population 2 presented in Tables 5.2 and 5.3 show the average and range of agreement across all mathematics and science free-response items, for each country. These results, showing the percentage of exact agreement for both the correctness score and the full two-digit diagnostic code, reveal a high degree of agreement for the countries that documented the reliability of their coding. Exact agreement between the first and second independent coders was particularly high for the first digit of the code, indicating the correctness score given the response. Since achievement on the TIMSS tests was estimated using only this first digit, it seems reasonable to conclude that scorer agreement within countries was robust. It appears that the use of open-ended items did not lower the reliability of the TIMSS tests, at least from a within-country perspective for countries providing within-country reliability data. Unfortunately, lack of resources precluded several countries from providing this information.

Naturally, the goal was to have 100% or perfect agreement between coders. In actuality, agreement above 85% is considered quite good, and above 70% acceptable. For the mathematics items, a very high percentage of exact agreement was observed for all countries, with averages across items for the correctness score ranging from 97% to 100%

and an overall average of 99% across all 26 countries. In addition, all countries had at least 77% agreement on all items. While the percentage of exact agreement for science items was somewhat lower than for the mathematics items, it is still quite good, with averages across items for the correctness score ranging from 88% to 100% and an overall average across the 26 countries of 95%. Also, nearly all countries had greater than 70% agreement on all items. Percentages of agreement below 70% may be a cause for concern. In fact, as part of the database review prior to scaling the TIMSS achievement data, countries were alerted about items where scoring agreement was below 70%. In several instances, this information uncovered a misunderstanding in the coding approach and the student responses were recoded before the achievement data were scaled.

Although the results in Tables 5.2 and 5.3 indicate a high degree of within-country coder agreement in assigning the overall score to students' responses, the data indicate less agreement concerning the second coding digit, designed to provide a more detailed view of the type of response. Even for the second digit, however, agreement was quite respectable, with averages across items ranging from 89% to 99% for mathematics items and from 73% to 98% for science items. Nevertheless, depending on the items and countries involved, some care should be taken in making comparisons across countries at this finer level.

Table 5.2
TIMSS Within-Country Free-Response Coding Reliability Data for Mathematics Items*

Country	Correctness Score Agreement			Diagnostic Code Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	98%	90%	100%	90%	61%	98%
Belgium (Fl)	100%	98%	100%	99%	92%	100%
Bulgaria	98%	93%	100%	94%	59%	100%
Canada	98%	85%	100%	92%	70%	99%
Colombia	99%	97%	100%	96%	91%	100%
Czech Republic	98%	77%	100%	95%	68%	100%
England	100%	96%	100%	97%	89%	100%
France	100%	96%	100%	98%	93%	100%
Germany	98%	89%	100%	94%	75%	100%
Hong Kong	99%	94%	100%	96%	84%	100%
Iceland	98%	84%	100%	91%	73%	100%
Iran, Islamic Rep.	98%	94%	100%	93%	70%	100%
Ireland	99%	95%	100%	97%	83%	100%
Japan	100%	96%	100%	99%	90%	100%
Netherlands	98%	87%	100%	91%	68%	100%
New Zealand	99%	95%	100%	95%	81%	100%
Norway	99%	90%	100%	95%	79%	100%
Portugal	98%	88%	100%	93%	82%	99%
Russian Federation	99%	94%	100%	96%	84%	100%
Scotland	97%	81%	100%	89%	63%	99%
Singapore	99%	95%	100%	98%	87%	100%
Slovak Republic	97%	84%	100%	91%	70%	98%
Spain	98%	88%	100%	94%	75%	100%
Sweden	99%	90%	100%	94%	75%	100%
Switzerland	100%	95%	100%	98%	83%	100%
United States	99%	95%	100%	96%	85%	99%
AVERAGE	99%	91%	100%	95%	78%	100%

*Based on 26 mathematics items, including 6 multiple-part items.

Note: Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and range

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 5.3
TIMSS Within-Country Free-Response Coding Reliability Data for Science Items*

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	91%	69%	99%	78%	48%	97%
Belgium (Fl)	100%	95%	100%	98%	82%	100%
Bulgaria	91%	63%	100%	81%	50%	100%
Canada	92%	76%	100%	80%	59%	99%
Colombia	97%	83%	100%	91%	73%	100%
Czech Republic	96%	87%	100%	90%	61%	100%
England	97%	90%	100%	91%	65%	100%
France	99%	95%	100%	97%	89%	100%
Germany	94%	81%	100%	84%	66%	100%
Hong Kong	94%	72%	100%	87%	56%	100%
Iceland	95%	74%	100%	83%	22%	98%
Iran, Islamic Rep.	88%	67%	100%	73%	33%	99%
Ireland	95%	87%	100%	89%	69%	100%
Japan	100%	96%	100%	98%	87%	100%
Netherlands	92%	75%	100%	79%	17%	100%
New Zealand	97%	90%	100%	90%	63%	100%
Norway	95%	87%	100%	91%	71%	100%
Portugal	96%	88%	100%	91%	75%	100%
Russian Federation	96%	87%	100%	91%	73%	100%
Scotland	89%	73%	99%	74%	52%	96%
Singapore	98%	92%	100%	95%	86%	100%
Slovak Republic	92%	62%	100%	81%	43%	100%
Spain	95%	85%	100%	88%	73%	98%
Sweden	94%	80%	100%	83%	54%	99%
Switzerland	98%	93%	100%	93%	85%	99%
United States	97%	90%	100%	89%	74%	100%
AVERAGE	95%	82%	100%	87%	63%	99%

*Based on 33 science items, including 4 multiple-part items.

Note: Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and ranges.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

5.4 IMPLEMENTING THE CROSS-COUNTRY RELIABILITY STUDY

At the Salzburg National Research Coordinators' meeting in late 1994, the NRCs suggested that TIMSS also should obtain information about coding reliability across countries. Fortunately, the additional funds for the quality assurance program included some resources for this purpose, and staff set out to design and implement such a study. The TIMSS International Coding Reliability Study was conducted in December 1995 in Boston.

Considering the schedule for TIMSS and its resources, it was clear from the outset that the cross-country coding study could be ambitious, but would be far from all-inclusive. Thus, the purpose would need to be focused, and choices would need to be made about which TIMSS populations to include in the study, numbers of items, and so forth. The next sections discuss the issues involved, the decisions made, and the procedures used in conducting the study.

5.4.1 OVERALL PURPOSE

The goal of the study was to document the level of agreement across countries in coding student responses to the mathematics and science free-response items. More complex aims were discussed, such as studying the sources of potential bias among countries and obtaining a sense of how differences in free-response coding might have affected the overall scores for countries. The TIMSS Technical Advisory Committee, in particular, supported these more complex goals. However, these aims remained largely beyond operational feasibility given the TIMSS schedule and budget.

5.4.2 POPULATION

Free-response items played a substantial role in all of the TIMSS tests, including mathematics and science at Populations 1 and 2. At Population 3, there were three possibilities–one test for the general population, a second for the subpopulation having studied advanced mathematics, and a third for the subpopulation having studied physics. Valuable information could have been obtained from studying the scoring reliability of many of the items included in these various tests. However, since Population 2 was the only mandatory population for participation in TIMSS, it was the population selected for the cross-country study of coding reliability.

5.4.3 NUMBER OF ITEMS

At Population 2, TIMSS included 26 free-response items in mathematics and 33 in science. Clearly, TIMSS would have preferred a reliability study involving all of these items, but coders from the participating countries simply could not commit to a study of such extensive proportions. After careful consideration, three of the eight books at Population 2 were considered appropriate as a basis for the study. Together, these three books included

15 mathematics items and 18 science items. One mathematics and one science item were eliminated to better balance the workload for the coders. As shown in Table 5.4, 31 items were involved in the reliability study–14 mathematics items and 17 science items. The study included about half of all of the free-response items at Population 2 – 54% of the mathematics items and 52% of the science items.

Table 5.4
Number of Mathematics and Science Items in Cross-Country Coding Reliability Study

Booklet	Mathematics			Science			Mathematics and Science Combined		
	SA	ER	Total	SA	ER	Total	SA	ER	Total
#3	2	4	6	2	0	2	4	4	8
#7	3	1	4	4	2	6	7	3	10
#8	4	0	4	8	1	9	12	1	13
Total	9	5	14	14	3	17	23	8	31

Notes: SA=Short Answer; ER=Extended Response

5.4.4 NUMBER OF STUDENT RESPONSES PER ITEM

Three considerations emerged in making decisions about how many student responses to each item should be included in the study:

- The ability to link to the within-country reliability studies
- The language of testing
- The overall coding burden.

To strengthen the study, it was based on the same student responses as the within-country reliability studies. Again, the preference was to involve the full reliability sample across the selected items for all of the participating TIMSS countries. However, the overall coding burden led to a final decision to use half-samples.

The many different languages of testing were one of the most difficult obstacles in conducting the study. The original notion involved student responses from all participating countries. After collecting information about the availability of bilingual coders, however, it became clear that trying to incorporate student responses in a number of languages into the study (e.g., German, French, Spanish, and the Scandinavian languages as well as English) simply was not going to be feasible. To provide information about the coding in the TIMSS countries, the study needed to involve the actual coders from the participating countries. A number of these coders were fluent in English as well as their own language, but very few were bilingual or multilingual in the other languages of interest. On the other hand, there was consensus that translating the students' responses into English introduces unknowns as well as being expensive and time-consuming. Under the circumstances, the best approach appeared to be using responses provided in English for the study. This enabled the use of original student responses and permitted participation by all countries wishing to work on the study.

Once the decision was made to base the study on student responses in English, 7 countries that administered the test in English were asked to provide student responses from their TIMSS testing at Population 2. Each country was asked to provide 50 student responses for each of the 31 items, essentially drawn from every other booklet in their within-country reliability samples. Since there were 7 English-test countries each supplying 50 responses for each item, the coding study involved 350 responses to each of the 31 items. This procedure resulted in a corpus of 10,850 student responses to serve as the basis of the study. The 7 countries devoting time and energy to supplying student responses included Australia, Canada, England, Ireland, New Zealand, Singapore, and the United States.

5.4.5 NUMBER OF PARTICIPATING COUNTRIES AND CODERS

A total of 39 coders from 21 countries participated in the international reliability study. Participation was voluntary, and all countries were invited to participate. Table 5.5 lists the countries that participated, and the names of the coders are listed in Appendix G. Countries could send as many as two coders, and all of the countries participating in the study did so except Canada, France, and Germany (they each sent one coder). Two coders per country enabled the study to be conducted in one week. It also enabled countries that had divided responsibility for the coding task by subject area to send one coder who specialized in science and another who specialized in mathematics.

Table 5.5
Countries Participating in Cross-Country Reliability Study

Australia	Ireland	Romania
Bulgaria	Latvia	Russian Federation
Canada	Lithuania	Singapore
England	New Zealand	Slovak Republic
France	Norway	Sweden
Germany	Philippines	Switzerland
Hong Kong	Portugal	United States

5.4.6 TWO GROUPS OF ITEMS AND CODERS

In order to accomplish all of the coding involved in the study during one week, the 31 items were divided into sets of 15 and 16 items. The division was essentially according to mathematics and science items, but because the science items take more time to code there also was an attempt to balance the workload between the two groups. Item Set 1 contained 12 mathematics items and 4 science items; Item Set 2 contained 13 science items and 2 mathematics items.

The coders also were divided into two groups, with one coder from each country in each of the groups. Information about the division of items was sent to the countries and coders in advance so that coders could receive refresher training in the items they were to score. Coders were to bring their own coding guides so that they could follow as closely as possible the procedures used in the within-country scoring. For Canada, France, and Germany (the three countries with only one coder), the coders elected to score Item Set 1. Thus, 21 coders worked on scoring Item Set 1 and 18 on scoring Item Set 2.

Because time permitted, 4 mathematics and 8 science items were scored by both groups of coders. Although this was not part of the original plan, it provides an important link between the two groups of coders. During debriefing at the end of the study, it was ascertained that for the countries participating the study, coder responsibilities at Population 2 were more likely to have been assigned by booklet than by subject area. Of the study participants, only the Russian Federation and Hong Kong specialized by subject area. Even though most coders had backgrounds predominantly in either mathematics or science, during the actual coding in their countries they had scored both mathematics and science questions.

5.4.7 THE DESIGN FOR EACH GROUP OF CODERS

As shown in Table 5.6, the 350 student responses for each item were divided into seven stacks of 50 responses. These stacks included responses across all seven countries supplying student responses, with each stack containing seven to eight responses from each of the countries. The responses for each item were organized to be distributed across coders according to a balanced rotated design. The seven stacks were placed into groups of three, such that every stack appeared with every other stack. Also, in this assembly care was taken that each stack appeared once as the first set of student responses, once as the second set, and once as the third set. Each coder, then, scored three stacks of responses for each item. This meant that for each item, each coder scored a total of 150 student responses (comprising 21 to 22 responses for each of the 7 English-test countries). This design also ensured that every coder shared a stack of at least 50 student responses with every other coder scoring the same set of items.

Table 5.6
The Design for Assigning Student Responses to Coders

Coder	Stacks	Each Stack
Coder A	1, 7, 5	• 50 Student Responses • Responses from all 7 countries - 8 responses from one country - 7 responses from the other 6 countries
Coder B	2, 1, 6	
Coder C	3, 2, 7	
Coder D	4, 3, 1	
Coder E	5, 4, 2	
Coder F	6, 5, 3	
Coder G	7, 6, 4	

Given that the design for assigning student responses to coders yielded seven combinations of the three stacks of student responses, and that the study involved 21 coders scoring Item Set 1 (primarily mathematics), there were three full rotations of coders for Item Set 1. Since for each rotation the combination of stacks already ensured that each stack and each student response in it was coded by three coders, the three rotations resulted in each student response being scored by coders from nine different countries (including one from the country that did the original coding). A similar situation existed for Item Set 2, where 18 coders participated. Here, though, there were not quite enough coders for three full rotations. For Item Set 2, not all responses were scored by nine coders, some receiving seven or eight codes depending on the rotation. For the 12 items scored by both groups of coders, student responses received 16 to 18 codes.

5.4.8 OPERATIONAL ASPECTS OF THE STUDY

The TIMSS International Study Center prepared the necessary set of student responses for each coder participating in the study. Thus, within daily guidelines specifying which three to five items were to be scored each day, each coder was able to work at his or her own pace and the International Study Center could rest assured that the coding sequences were being maintained in accordance with the study design. The International Study Center engaged two supervisors from the United States TIMSS free-response coding effort to act as the table leaders for the two groups. They began each coding session by giving their group of coders an overview of the work for the day. Then, each group of coders began with a particular item. Within the group, in accordance with the design shown in Table 5.6, each coder had possession of the 150 responses they were to score for that item. Once coders had finished coding those responses, they moved on the next item until the day's work was completed.

Codes for each stack of 50 responses were recorded on answer sheets devised by the International Study Center that included the booklet and item number, the country and student identification numbers for the responses, the coder's identification, and the codes given to each response. After scoring an item, the coder submitted the answer sheets to a clerk so they could be checked for completeness and to ensure that the codes were within the range valid for the item. This quality control step was conducted throughout the study.

The data from the coding study were entered by a professional data entry agency. The entry process and the database were subjected to a series of quality control checks, including the accuracy of data entry and any appropriate recoding necessitated by the use of special within-country codes by some coders.

5.5 THE RESULTS OF THE INTERNATIONAL CROSS-COUNTRY CODING STUDY

5.5.1 PERCENT AGREEMENT

To provide direct comparisons with the results obtained from the within-country reliability studies, the International Study Center computed the percentage exact agreement for both correctness scores and diagnostic codes. The entire student sample of 350 responses for each free-response item was used to compute these measures. All coder pairs who coded a common subset of at least one stack of 50 student responses contributed to the overall percent agreement measure for each item. In the cross-country study design, each student response was coded by 7 to 18 coders; the measure of percent agreement obtained reflects an overall pairwise percentage of agreement based on all possible coder pairs for each item. As a result of the study design, nearly all of the across-coder comparisons included in the percent agreement calculations are across-country comparisons. For the items coded by both groups of coders, the percent agreement measures include approximately 2% comparisons between coders from the same country.

Table 5.7 summarizes the average percent agreement across the 31 items used in the international reliability study and compares these results with the corresponding within-country measures for the same items. The within-country measures are reported as the average and the range of percent agreement measures across the 26 countries submitting within-country reliability results.

Table 5.7
Average Percent Exact Agreement for All Items in International Free-Response Coding Reliability Study

Subject Area	Number of Items[1]	Average Correctness Score Agreement				Average Diagnostic Code Agreement			
		International Study	Within-Country Study[2]			International Study	Within-Country Study[2]		
			Average	Min	Max		Average	Min	Max
Mathematics	14	97%	98%	92%	100%	89%	93%	83%	99%
Science	17	87%	95%	82%	100%	71%	86%	63%	99%
OVERALL AVERAGE	31	92%	96%	87%	100%	80%	90%	73%	99%

[1] Includes four math and one science 2-part items. Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and ranges.

[2] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

A high overall average exact agreement of 92% for correctness scores and 80% for diagnostic codes was obtained for the 31 items in the international study. These are somewhat lower than the corresponding average within-country results of 96% and 90%. For the mathematics items, high average percent agreements of 97% for correctness score and 89% for diagnostic code were obtained, which compare very favorably with the respective average within-country measures of 98% and 93%. The science items, which in general use more complex coding rubrics than the mathematics items, had average percent agreement values that were lower for both the international and within-country studies. In addition, the difference between the two studies was greater for the science items, with average cross-country percent agreement measures of 87% and 71% compared with within-country measures of 95% and 86% for diagnostic code and correctness score, respectively. Although the average international study results are lower than the corresponding within-country measures, they still fall well within the range of within-country results obtained across all 26 countries. Moreover, they exceed by a substantial margin the correctness score agreement threshold of 70% used to identify items exhibiting within-country coding reliability problems.

More detail about the percent agreement measures for each item in the international study is shown in Tables 5.8 and 5.9 for mathematics items and science items, respectively. These tables also show the total number of individual comparisons used in computing the cross-country percent agreement measures, since this number varies for different items due to the rotation design and the division of items and coders into two sets of unequal size. The percent agreement for the twelve items that were coded by both groups of coders was first computed for the two sets separately. A comparison of the two measures revealed a maximum difference of less than 5% for any item. The differences in most cases reflected a slightly higher percent agreement for the group to which the item was originally assigned. Since the differences between coder groups was small, the calculations for these twelve items include comparisons for coders in both groups to obtain the broadest across-country comparisons possible.

The percent of exact agreement for each mathematics item was very high, with only two items having correctness score agreement measures below 90%. Diagnostic code agreements were, in general, lower, ranging from 61% to 98%. Even for the diagnostic agreement, however, 13 out of 18 items had percent agreement greater than 90%. For the majority of items, the difference between the international and within-country average diagnostic code percent agreement measures was 5% or less, and for all but two items, the international measure fell within the range of within-country values. For the correctness score agreement, all items were well within the range of the within-country results. The percent of exact agreement for science items was, in general, lower and exhibited a much broader range, with diagnostic code agreement ranging from 50% to 98%. When only the first-digit correctness score is considered, a percent exact agreement range of 72% to 99% is observed. Even the items with the lowest international diagnostic agreement have correctness score agreement levels that fall within the range of within-country measures and exceed the 70% threshold. Also, only a few of the science items had large diagnostic code agreement differences between the international and within-country measures (20% or greater).

Table 5.8
Percent Exact Agreement for Coding of Mathematics Items

Item Label[1]	Total Valid Comparisons[2]	Correctness Score Agreement				Diagnostic Code Agreement			
		International Study	Within-Country Study[3]			International Study	Within-Country Study[3]		
			Average	Min	Max		Average	Min	Max
M1	9150	100%	99%	96%	100%	97%	97%	84%	100%
[4] M2A	46050	100%	100%	96%	100%	98%	98%	94%	100%
M3	12600	99%	99%	95%	100%	98%	97%	92%	100%
M4	46050	99%	99%	96%	100%	99%	98%	87%	100%
M5	45985	99%	100%	96%	100%	97%	98%	92%	100%
M6	12600	99%	99%	98%	100%	97%	98%	91%	100%
M7	12600	99%	99%	96%	100%	95%	98%	92%	100%
M8	12600	99%	99%	94%	100%	91%	95%	89%	100%
M9	9150	99%	99%	94%	100%	94%	97%	90%	100%
[4] M2B	46050	99%	99%	95%	100%	91%	94%	74%	100%
[4] M10A	45938	98%	100%	98%	100%	95%	97%	90%	100%
[4] M11A	12592	97%	98%	84%	100%	91%	94%	77%	100%
M12	12600	97%	99%	95%	100%	93%	95%	88%	99%
[4] M11B	12600	96%	98%	95%	100%	74%	88%	68%	100%
[4] M13A	12600	95%	97%	90%	100%	85%	92%	75%	99%
M14	12600	91%	96%	81%	100%	77%	89%	72%	98%
[4] M13B	12592	89%	96%	84%	100%	71%	88%	75%	100%
[4] M10B	46050	84%	93%	77%	99%	61%	82%	61%	97%
AVERAGE MATH ITEMS		97%	98%	92%	100%	89%	94%	83%	100%

[1] See Appendix H for item descriptions and coding guides.
[2] Values for items coded by the same coder group differ slightly due to a small number of missing responses or invalid codes.
[3] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.
[4] Two-part items; each part is analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 5.9
Percent Exact Agreement for Coding of Science Items

Item Label[1]	Total Valid Comparisons[2]	Correctness Score Agreement				Diagnostic Code Agreement			
		International Study	Within-Country Study[3]			International Study	Within-Country Study[3]		
			Average	Min	Max		Average	Min	Max
S1	9078	99%	99%	95%	100%	98%	97%	80%	100%
S2	46035	94%	97%	77%	100%	74%	86%	64%	100%
S3	9150	93%	96%	81%	100%	85%	91%	54%	100%
S4	12600	93%	95%	83%	100%	67%	80%	52%	99%
S5	46050	92%	97%	88%	100%	78%	88%	58%	100%
S6	46050	91%	96%	90%	100%	79%	91%	79%	99%
[4] S7A	9150	90%	95%	83%	100%	71%	87%	67%	99%
[4] S7B	9150	89%	95%	87%	100%	77%	89%	74%	98%
S8	45930	89%	96%	90%	100%	70%	84%	65%	98%
S9	46050	88%	93%	74%	100%	74%	87%	64%	100%
S10	9150	88%	96%	86%	100%	83%	91%	65%	100%
S11	9122	86%	95%	86%	100%	72%	87%	61%	100%
S12	45930	86%	95%	81%	100%	59%	80%	53%	96%
S13	46034	82%	93%	74%	100%	66%	87%	65%	100%
S14	9150	80%	93%	82%	100%	59%	82%	47%	100%
S15	46050	78%	92%	75%	100%	70%	89%	69%	99%
S16	12600	75%	91%	74%	100%	51%	78%	55%	100%
S17	9129	72%	90%	70%	100%	50%	82%	59%	100%
AVERAGE SCIENCE ITEMS		86%	94%	81%	100%	70%	86%	62%	99%

[1] See Appendix H for item descriptions and coding guides.
[2] Values for items coded by the same coder group differ slightly due to a small number of missing responses or invalid codes.
[3] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.
[4] Two-part items; each part is analyzed separately

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

5.5.2 INVESTIGATING DIFFERENCES BETWEEN THE CROSS-COUNTRY AND WITHIN-COUNTRY STUDIES

Although the cross-country percent agreement measures are systematically somewhat lower than those reported from the within-country samples, it should be remembered that these results also reflect differences between the two types of studies. First, for many coders the student responses used in the international study were in a different language and reflected a different culture from those encountered by the coders in their own country. Both of these factors added to the complexity of coding in the cross-country study. Second, while the within-country measures were obtained during the actual coding sessions in each country, when the training and guides were fresh in the coders' minds, for most participants the international study was conducted several months after the main study coding sessions. Although each coder involved in the international study received some refresher training before participation, the potential for coder agreement might well have decreased after the main study coding sessions. Third, the coding environment in the international study was somewhat different from those within the individual countries. For example, several countries indicated that during their country's coding sessions, coding difficulties encountered with some student responses were resolved by consensus, which was not the case in the international study. Also, coders in the international study had to work with photocopies rather than the original booklets.

To investigate differences between the percentage of agreement reported for the two types of reliability studies, the 12 items that were coded by both groups of coders were used to determine the percent diagnostic code agreement between the two coders from each country. The average of these measures for the 18 countries that sent two coders to the international study was used as a measure of the percent agreement obtained under the conditions of the international study, excluding any across-country coder effects. Table 5.10 provides a comparison of this within-country measure from the international study with both the across-country measure and the average percent agreement from the within-country reliability studies conducted in 26 countries.

Table 5.10
Comparison of Diagnostic Code Agreement from the International and Within-Country Studies

	International Study		Within-Country Studies Average[3]
	Average Within-Country Percent Agreement[1]	Overall Across-Country Percent Agreement[2]	
Mathematics Items			
M2A	98%	98%	98%
M2B	91%	91%	94%
M4	99%	99%	98%
M5	98%	97%	98%
M10A	94%	95%	97%
M10B	60%	61%	82%
Mathematics Average	90%	90%	95%
Science Items			
S2	76%	74%	86%
S5	80%	78%	88%
S6	81%	79%	91%
S8	70%	70%	84%
S9	77%	74%	87%
S12	62%	59%	80%
S13	68%	66%	87%
S15	70%	70%	89%
Science Average	73%	72%	86%
Overall Average	80%	80%	90%

[1] Average of percent agreement between the two coders from each country for 18 of the countries in the international study in Table 5.5.

[2] Percent agreement from the international study based on all coder comparisons.

[3] Average of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

These results show that the within-country and across-country percent agreement measures from the international study are comparable for all items, and that both are lower than the corresponding measure from the within-country study. The systematically lower percent agreement of the international study thus appears to be due primarily to situational and contextual differences in the way the two measures were obtained rather than to decreased across-country coding reliability. Based on these results, the reliability of the international free-response coding from the main study coding sessions would be expected to be no lower than that from the within-country results, and therefore, should be quite good for most items.

5.5.3 COMPARING CODE AGREEMENT FOR INDIVIDUAL COUNTRIES

Another measure of across-country agreement is to compare how well the coders from each of the participating countries agree with the other coders in the same group. Table 5.11 presents the diagnostic code percent exact agreement between each participating coder and the coders from other countries in the same group.

These results, averaged across all items in the item set, reveal that there is a good level of consensus within each group of coders and that the agreement for individual coders is quite comparable across countries. For Item Set 1, the individual coder agreement ranges from 77% to 85% with an average of 82%. The agreement for Item Set 2 is somewhat lower, ranging from 71% to 80%, with an average of 77%. These differences in agreement within the two sets of coders, however, reflect the nature of the items assigned to them, with Item Set 1 being predominantly mathematics and Item Set 2 predominantly science, which were more complicated to code.

Table 5.11
Comparison of Exact Percent Diagnostic Code Agreement for Individual Countries[1]

Country	Item Set 1[2] 12 Math 4 Science	Item Set 2[3] 13 Science 2 Math
Australia	84%	77%
Bulgaria	79%	76%
Canada	84%	*
England	84%	78%
France	84%	*
Germany	82%	*
Hong Kong	83%	75%
Ireland	83%	76%
Lithuania	79%	79%
Latvia (LSS)	82%	75%
Norway	84%	80%
New Zealand	83%	80%
Philippines	77%	76%
Portugal	83%	74%
Romania	83%	76%
Russian Federation	79%	76%
Slovak Republic	80%	71%
Singapore	83%	79%
Sweden	81%	77%
Switzerland	79%	78%
United States	85%	79%
AVERAGE	82%	77%

[1] Percent agreement between each coder and the coders from all other countries averaged over all items in the item set.

[2] Items in Set 1: M2, M3, M4, M5, M6, M7, M8, M10, M11, M12, M13, M14, S2, S4, S13, S16

[3] Items in Set 2: M1, M9, S1, S3, S5, S6, S7, S8, S9, S10, S11, S12, S14, S15, S17

*No Item Set 2 coders from Canada, France and Germany.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

5.5.4 FREQUENCIES OF DIAGNOSTIC CODE AGREEMENT

Contingency tables showing the cumulative frequencies of pairwise code combinations were computed for all items in the international study. These tables can be used to obtain detailed information about the nature of code discrepancies. Contingency tables and coding guides for all items in the international study are included in Appendix H. An example contingency table and the corresponding item and coding guide are shown in Figure 5.1 for item S11. This item was one for which the percent diagnostic code agreement was moderate, indicating that a number of code discrepancies are expected.

The cell counts in the contingency table in Figure 5.1 indicate the total number of times, over the entire set of student responses, that for the same student response one coder in a pairwise comparison gave the code corresponding to the row position, while another coder gave the code corresponding to the column position. The simple percent exact agreement for both the diagnostic code and the correctness score may be computed from these frequencies of nominal agreement. The diagnostic code percent exact agreement is computed from the sum of the diagonal cells, where the two paired codes match exactly, while the correctness score percent exact agreement is computed from the sum of all the shaded cells, where the first digits of the paired codes correspond to the same correctness score.

The TIMSS free-response coding guides are specific to each item; the coding guide for item S11 is shown as an example. This coding guide has 8 valid diagnostic codes, 3 that correspond to a correct response (first digit of 1) and 5 that correspond to an incorrect response (3 codes with a first digit of 7 for an incorrect response and 2 codes with a first digit of 9 for a nonresponse). A second digit of 9 (code 19 or 79) is used for responses that are judged to be within the level of correctness indicated by the first digit but do not fit any of the other specific diagnostic codes. The level of disagreement about specific diagnostic codes varies substantially from item to item, but there are some patterns of code disagreement that are common to many of the items. Some of these types of patterns can be observed with the S11 example item.

Item S11 had a diagnostic code percent agreement of 72% and a correctness score percent agreement of 86%. The frequencies of matched codes indicate that approximately 10% of the code comparisons reflect diagnostic code disagreements where two paired codes on a student response are either both correct or both incorrect, but only one coder used a specific diagnostic code (10,11 or 70,71), while the other used an Other code (19 for Other Correct or 79 for Other Incorrect). Another 3% of code discrepancies are due to student responses that were coded as correct but where there was disagreement on whether the 10 or the 11 diagnostic code was given. The most common code discrepancies contributing to the lack of correctness score agreement, approximately 12%, is due to student responses that one coder scored as correct (10, 11, or 19), while another coder gave a code of Other Incorrect (79). This types of code discrepancy was found to be fairly common across many of the items investigated in the international study, with the use of the Other codes accounting for a substantial portion of disagreement in both the diagnostic code and the correctness score. These code disagreements did not usually result in low overall correctness score agreement, however. Another type of discrepancy that can reduce diagnostic code agreement is the interpretation of what constitutes a nonresponse (code 90 or 99). Although the 99 code was to have been reserved for absolutely blank responses, sometimes very brief partial responses also were given a code of 99. In most instances, however, the disagreements were between the 90 and the 79 codes. The extent of that disagreement is understandable given that both codes reflect types of incorrect responses that can be difficult to interpret. For all items, less than 3% of code comparisons reflected disagreements involving the 90 or 99 codes.

Figure 5.1
Item Description, Frequencies of Diagnostic Code Agreement and Coding Guide for Example Item S11

Carbon Dioxide Fire Extinguishers

Item S11

Carbon dioxide is the active material in some fire extinguishers. How does carbon dioxide extinguish a fire?

	FREQUENCIES OF MATCHED CODES								
	10	11	19	70	71	79	90	99	ROW SUM
10	3508	297	340	2	42	542	18	0	4749
11		490	116	29	59	365	1	0	1060
19			43	3	16	165	4	0	231
70				296	10	126	3	0	435
71					265	286	10	0	561
79						1086	89	0	1175
90							111	36	147
99								764	764

TOTAL VALID COMPARISONS	9122

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that carbon dioxide keeps oxygen away; response includes explicit reference to oxygen.
11	Mentions that carbon dioxide keeps "air" away.
19	Other correct.
Incorrect Response	
70	Mentions that carbon dioxide cools down the fire.
71	Refers to a material in carbon dioxide.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

5.5.5 GENERALIZABILITY OF FREE-RESPONSE ITEM SCORES

An analysis of variance of the international reliability study data was used to estimate generalizability coefficients for both the country-level average scores and the student-level scores on each of the free-response items in the international study. The generalizability coefficients computed are a measure of the reliability of the free-response item scores in that they reflect the proportion of observed variance due to true score variance for the object of measurement. In the computation of generalizability coefficients, specific sources of error variance are identified according to the design of the study and the definition of the object of measurement. Generalizability coefficients computed for each of the items are shown in Table 5.12 for mathematics items and Table 5.13 for science items.

For the country-level averages, the object of measurement is the average score for each of the seven English-test countries contributing student responses. The relative error variance has contributions from both the variance due to students within countries and the variance due to rater effects (both main and interaction effects). The generalizability coefficient reflects the reliability of the relative ranking of a country's average score on an item based on the total sample of students, given that each student response receives one rating by a rater within that country. In general, there were many raters participating in coding, and the full set of student responses in each country was divided among these raters. Therefore, the generalizability coefficient is a function of the total sample size and the total number of raters involved in rating the entire set of student responses in each country. Generalizability for all items increases as each of these levels increases, but for the typical sample sizes used in the TIMSS study, generalizability is more sensitive to the number of raters than to increases in sample size for many items. The sensitivity of generalizability to numbers of raters and students differs from item to item, depending on the relative contribution to total variance due to country, student, and rater effects.

In Tables 5.12 and 5.13, generalizability coefficients are presented for two sample sizes (500 and 1000) and three levels of number of raters (5, 15 and 25) to be representative of the ranges of values encountered in most of the countries in the TIMSS study. The generalizability of country-level averages is quite high for most of the mathematics and science items, with generalizability coefficients greater than 0.7 at the lower levels of raters and students for all but three of the science items and all but one of the mathematics items. Increasing the number of raters from 5 to 15 results in an increase in the generalizability to above 0.7 for all of these items. This analysis suggests that the generalizability of country-level averages on free-response items would be an issue only if very small numbers of raters were involved in the coding in each country. Also, since the generalizability analyses reflect only the seven English-test countries represented in the international study, the variance in average scores for this particular set of countries is lower than what would be obtained if all TIMSS countries were represented in the analysis. Provided that the rater and student effects are comparable for the countries not included in the generalizability study sample, it is likely that the generalizability coefficients presented here underestimate the generalizability of country-level averages for the entire TIMSS population.

Table 5.12

Generalizability of Scores on Free-Response Mathematics Items Based on the International Reliability Study Sample

Item	Generalizability Coefficients for Country-Level Averages[1]						Generalizability Coefficient for Student-Level Scores[4]
	Sample Size = 500[2]			Sample Size = 1000[2]			
	Number of Raters[3]			Number of Raters[3]			
	5	15	25	5	15	25	
M8	0.99	0.99	0.99	1.00	1.00	1.00	0.98
M1	0.99	0.99	0.99	1.00	1.00	1.00	0.99
M5	0.99	0.99	0.99	0.99	0.99	0.99	0.99
M9	0.99	0.99	0.99	0.99	0.99	0.99	0.98
M3	0.99	0.99	0.99	0.99	0.99	0.99	0.98
M6	0.99	0.99	0.99	0.99	0.99	0.99	0.99
[5] M11B	0.98	0.99	0.99	0.99	0.99	0.99	0.91
[5] M13B	0.98	0.99	0.99	0.99	0.99	0.99	0.85
[5] M11A	0.98	0.98	0.98	0.99	0.99	0.99	0.98
[5] M13A	0.97	0.97	0.98	0.98	0.99	0.99	0.97
M4	0.97	0.97	0.97	0.98	0.98	0.98	0.98
M12	0.97	0.97	0.97	0.98	0.98	0.98	0.92
M14	0.96	0.96	0.96	0.98	0.98	0.98	0.96
[5] M2A	0.95	0.95	0.95	0.97	0.97	0.97	0.99
M7	0.93	0.93	0.93	0.96	0.96	0.96	0.99
[5] M2B	0.92	0.93	0.93	0.95	0.96	0.96	0.95
[5] M10A	0.86	0.87	0.87	0.92	0.93	0.93	0.97
[5] M10B	0.58	0.74	0.79	0.61	0.79	0.84	0.69
Average	0.94	0.95	0.96	0.96	0.97	0.98	0.95

[1] Generalizability of the average country-level score on an item, based on one rating for each student.
[2] Total number of students within a country responding to each item.
[3] Total number of raters within each country scoring a subset of the student responses for each item.
[4] Generalizability of an individual student's score on an item, based on one rating.
[5] Two-part items; each part analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

BEST COPY AVAILABLE

Table 5.13
Generalizability of Scores on Free-Response Science Items Based on the International Reliability Study Sample

Item	Generalizability Coefficients for Country-Level Averages[1]						Generalizability Coefficient for Student-Level Scores[4]
	Sample Size = 500[2]			Sample Size = 1000[2]			
	Number of Raters[3]			Number of Raters[3]			
	5	15	25	5	15	25	
S9	0.97	0.98	0.98	0.98	0.99	0.99	0.90
S10	0.94	0.95	0.96	0.95	0.97	0.98	0.89
S17	0.93	0.97	0.97	0.94	0.97	0.98	0.66
S3	0.93	0.94	0.94	0.96	0.97	0.97	0.94
S6	0.92	0.95	0.96	0.94	0.97	0.97	0.82
S11	0.92	0.94	0.94	0.94	0.96	0.97	0.70
S2	0.90	0.92	0.93	0.93	0.95	0.96	0.86
S12	0.89	0.92	0.93	0.91	0.95	0.96	0.74
S4	0.88	0.93	0.94	0.90	0.95	0.96	0.54
[5] S7B	0.88	0.92	0.93	0.90	0.95	0.96	0.78
S1	0.87	0.87	0.87	0.93	0.93	0.93	0.99
[5] S7A	0.86	0.91	0.93	0.88	0.94	0.95	0.46
S8	0.84	0.89	0.90	0.87	0.93	0.94	0.80
S15	0.82	0.88	0.90	0.85	0.92	0.93	0.84
S14	0.76	0.87	0.89	0.78	0.89	0.92	0.59
S13	0.68	0.83	0.86	0.70	0.86	0.89	0.42
S16	0.60	0.79	0.84	0.61	0.81	0.87	0.56
S5	0.59	0.75	0.79	0.62	0.80	0.84	0.57
Average	0.84	0.90	0.91	0.87	0.93	0.94	0.72

[1] Generalizability of the average country-level score on an item, based on one rating for each student.
[2] Total number of students within a country responding to each item.
[3] Total number of raters within each country scoring a subset of the student responses for each item.
[4] Generalizability of an individual student's score on an item, based on one rating.
[5] Two-part items; each part analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

For the student-level scores, the object of measurement is the individual student score, and the relative error variance is due to the effect of the interaction between raters and students. The generalizability coefficient reflects the reliability of an individual student's score on an item, given that each student response receives only one rating. The person-by-rater interaction effect was found to vary substantially from item to item, particularly for the science items. The variance due to the person-by-rater interaction ranged from as low as 1% to as high as 50% of the total variance in student scores. This is reflected in the generalizability coefficients observed across the science items, which range from 0.42 to 0.99. Despite a low generalizability for a few items, for 11 out of 18 science items the student score generalizability was above 0.70. For mathematics items, the rater effects were much lower and the generalizability of the student scores was quite high for all but one of the items. Even for some of the science items with low individual score generalizability, however, the generalizability of the country-level averages was still quite high as it is based on a large number of student responses and raters. Since the goal of TIMSS is to report country-level averages and not individual scores, the lower generalizability for individual scores is not a concern for the international TIMSS reporting on the free-response items. These results serve as a caution, however, in performing secondary analyses that involve making any generalizations from individual student scores on specific items.

5.6 SUMMARY

Within resource constraints facing both the individual countries and the International Study Center, TIMSS has put considerable energy into the use of free-response items. Approximately one-third of the students' response time is devoted to free-response questions, which across the TIMSS tests encompasses about 300 free-response items. To provide diagnostic information about achievement, test development included an extensive effort to design scoring guides tailored to each of these questions. In the TIMSS two-digit scoring approach, the first digit indicates the correctness score (including levels of partial credit) and the second digit provides diagnostic information about the specific type of response.

Considering the number of items, number of countries, number of testing languages, and number of students involved, the scope of the free-response scoring effort was complex by anyone's standard. Therefore it was important for TIMSS to emphasize the importance of reliable scoring procedures. This includes very careful attention both to using reliable scoring procedures and to conducting studies to document the success of the procedures used.

Planning for TIMSS data collection included an ambitious series of training sessions for participating countries. The International Study Center conducted regional training sessions of essentially one week each to assist representatives of the national centers who would then be responsible for training personnel in their countries to apply the two-digit codes reliably. Nine sessions were held in total to accommodate participation by all countries in accordance with the different schedules in the Southern and Northern hemispheres. During the training sessions, participants were given detailed information about how to conduct free-response scoring and opportunities to practice the procedures, including substantial time in practicing scoring actual student responses according to the TIMSS guides.

The results from the within-country scoring reliability studies indicate that the percent of exact agreement among coders was very high, especially considering the many challenges underlying the effort. Each country was required to collect information about the reliability of its scoring procedures by having 10% of the student responses scored independently by two coders. Not all countries were able to afford this effort, but 26 countries provided data about the reliability of their scoring procedures. The average percent of exact agreement for the correctness score within each of the countries ranged from 97% to 100% on the mathematics items and from 88% to 100%. Average percentages of exact agreement for the diagnostic codes also were quite respectable, ranging from 89% to 99% for mathematics items and from 73% to 98% for science items.

The results of the international reliability study conducted using student responses from Population 2 also revealed a very high degree of across-country agreement among coders. Based on 350 student responses to each of 31 mathematics and science items, a total of 39 coders from 21 countries participated in the cross-country reliability study conducted by the International Study Center. The student responses used were randomly sampled from the within-country reliability samples of seven English-test countries: Australia, Canada, England, Ireland, New Zealand, Singapore, and the United States. A high overall average percentage of exact agreement of 92% for correctness scores and 80% for diagnostic codes was obtained.

In addition to documenting the high quality of the TIMSS free-responses scoring, various comparisons of the results from the TIMSS within-country and cross-country reliability studies reveal some interesting findings. Agreement was systematically higher for the mathematics items than for the science items. This seems reasonable, given that the coding guides for the mathematics items tended to be more straightforward. The results also indicate somewhat less agreement across countries than within countries, although further analyses reveal that these differences may be attributed primarily to differences in the conditions of the two types of studies. For example, for the international study many coders were not evaluating student responses in their native languages, so translation and cultural issues most likely made interpretation of responses more difficult. Also, for some coders several months had passed since the scoring effort in their own countries, and the coding task might not have been as familiar during the international study despite refresher training.

Generalizability coefficients computed for country-level averages and student-level scores indicate a high degree of reliability in the relative ranking of a country's average score based on using data from the TIMSS free-response items. The generalizability of country-level averages is quite high for most of the items, with coefficients generally greater than 0.7. As might be expected, the generalizability for an individual student's score on a particular item was found to be somewhat less stable for some items, ranging from 0.42 to 0.99. Since the goal of TIMSS is to report country-level and not individual-level results, the lower generalizability for individual scores is not a concern for reporting free-response item averages. In fact, all the TIMSS data from the reliability studies indicate that the scoring procedures were very robust both within and across countries. At least from the perspective of the quality of the free-response scoring, TIMSS can report the international achievement results with confidence.

REFERENCES

Mullis, I.V.S., Jones, C., and Garden, R.A. (1996). "Training Sessions for Free-Response Scoring and Administration of Performance Assessment" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc. Ref.: ICC918/NRC449). Chestnut Hill, MA: Boston College.

6. DATA CONSISTENCY CHECKING ACROSS COUNTRIES

Heiko Jungclaus
Michael Bruneforth

6.1 OVERVIEW

This section describes the data processing procedures used by the IEA Data Processing Center (DPC) in Hamburg. It describes the steps that were involved in cleaning the TIMSS data and standardizing the structure of the files across countries. It also describes the procedures that were implemented to facilitate the construction of the international database.

The TIMSS data processing procedures undertaken by the DPC had the following main objectives:

- To identify and document deviations from the international instruments and file formats
- To correct, whenever possible, identified errors in the data
- To make all changes to the data necessitated by inconsistent responses
- To provide detailed documentation on data quality, both at the country level and at the item or question level
- To create standardized file structures for the international data archive
- To provide countries with cleaned and weighted data
- To perform primary analysis to help countries review their data
- To implement modifications indicated by NRCs after their review of primary analysis.

6.2 DATA CLEANING

The main objective of data processing was to ensure the availability of clean data for further analyses. This process, hereafter called cleaning, included several steps. The main goals of cleaning were to identify, document, and, when necessary and possible, correct the following:

- Deviations from the international instruments (omitted questions or options, additional options)
- Deviations from the international default structure in the national file structures
- Systematic errors or deviations in data sets (e.g., deviating coding schemes)
- Formal inconsistencies within single observations (e.g., deviations from the hierarchical ID system, incorrect Test Indicator Variables)
- Problems in linking observations between files (e.g., teachers and students)

- Inconsistent tracking information between files (e.g., Grade ID at the student and teacher levels)
- Logical inconsistencies between the responses (e.g., inconsistencies between filter and dependent questions).

Although data cleaning focuses mainly on solving cleaning problems, a second important goal of data inspection was to identify and document insoluble problems as an indicator of data quality. Each country was provided with a set of national data documentation, which described any deviations from the international data structure, and summarized the results of the data cleaning for that country.

The cleaning described below has been performed for all data sets (up to ten data files per population).

6.2.1 STRUCTURE REPORT AND REVIEW OF NATIONAL DATA DOCUMENTATION

The first step after national TIMSS data were received was the inspection of the file structure. The national structure was automatically compared with the international default structure. The national structure database, if available, was compared with the international one. If countries did not send in their national electronic codebooks, some checks (e.g., identifying changes in coding schemes) could not be made.

The following deviations were noted:

- Missing variables (questions or items)
- Different variable lengths or number of decimals
- Different coding schemes
- Additional national variables
- Gang-punched variables.

After inspection of the incoming data, the national documentation (NRCs' reports on survey activities and Data Management Forms) was compared with the results of the structure check. Also, questionnaires were compared with the international default instruments. However, several countries sent in the tracking forms, instruments, and documentation which were incomplete. Correcting deviations and verifying procedures therefore required more communication with national centers than it otherwise would have. The results of the structure check and the review of national documentation and instruments are summarized in the section entitled *Report on File Structure and Systematic Deviations* in the national data documentation, which indicates:

- Whether or not international options were used
- The omission of items or questions that were part of the international core
- Changes in the coding schemes
- Changes in the identification variables

- Other problems.

6.2.2 PRECLEANING

After reviewing the NRCs' reports, additional documentation materials, the structure check, and the instruments, the DPC planned necessary changes to the data to solve all systematic problems. These adaptations had to be made before the automatic standard cleaning could be performed. This precleaning created data files that matched the standard structure given in the international codebook without losing additional country-specific information.

The process of precleaning was an individualized process depending on national deviations. For some countries, little or no precleaning was necessary; for others, complex programs had to be prepared.

The most frequent steps in precleaning were:

- Adjusting deviating identification systems to the international default (e.g., converting to the TIMSS standard all identification systems that did not provide a unique identification of cases or did not correctly indicate the class or school that a student attended)
- Adding dropped variables and recoding them to 'Not administered'
- Correcting data entry problems (e.g., replacing blanks with NA)
- Creating or completing missing information on teacher-student linkage in the student data
- Correctly coding 'Not administered' test booklets
- Adjusting national coding schemes to the international default (e.g., if countries used extra national options or options in a different order)
- Recreating missing international variables from national variables.

All changes made are documented in the *Report on File Structure and Systematic Deviations* in the national data documentation.

6.2.3 STANDARD CLEANING

When all data files matched the international structure exactly, the standard cleaning began. During standard cleaning, problems and deviations at the observation level were identified; in other words, standard cleaning focused on all deviations that applied to single respondents or groups of respondents (members of one class or teachers in one school).

When deviations were identified, the DPC compared them with the tracking forms, if available, and corrected them accordingly. Remaining problems were sent to the NRCs to be checked, and NRCs were asked to confirm obvious changes (e.g., deriving an ID from other IDs). In all cases in which no clear solution could be suggested, either by reviewing the documents or by asking the national centers, changes were made according to the cleaning rules.

For this phase, a new tool was developed. All deviations were recorded to a set of related databases. The *Cleaning Reports* in the national data documentation were generated automatically by the TIMSS Cleaning Program, which recorded the number and ID of the observation, the type of problem (indicated by a Problem Number), and the changes made to the database by the program. The program searched for over 160 different types of problems/deviations. For a detailed list of identified problems, refer to Appendix I.

All corrections that were undertaken were later identified by rerunning the same program over the cleaned file, so that the number and percentage of both corrected and unchanged problems could be checked. These changes were separated into automatic and manual changes. After the cleaning process has been finalized, this database can be used to prepare international problem statistics that would allow judgment of the data quality by country or by problem type.

The standard cleaning took place in several steps, beginning with the TIMSS Cleaning Program and continuing interactively with the national centers. Changes and corrections were made according to information from the tracking forms, information given by the NRCs, or the cleaning rules. All manual corrections were also archived in a separate database, which included the change and when it was made, and indicated a reason for the change.

6.2.4 CLEANING RULES

Most of the identified problems were solved in cooperation with national centers according to the national documentation. The national centers returned information on identified mispunches and corrected tracking information.

If the identified problems could not be clarified, cleaning rules had to be applied. The general idea behind the cleaning rules is explained in this section. The rules were applied to all cases in which a problem could not be solved (i.e. the respondent answered inconsistently) or the country did not respond to requests.

Some recoding was performed without asking the country. This occurred when decisions on 'Missing' versus 'Not checked' or 'Missing' versus 'Not administered' were necessary. Cases where respondents obviously did not follow the directions on the questionnaire were also corrected without asking for feedback from the NRCs.

Different rules had to be applied for different types of cleaning problems. In general, there were two types of variables and two ways to handle problems.

Variables of the first type were those containing formal information assigned or obtained by the National Center (i.e. not obtained from questionnaires). These included IDs, Test Indicator Variables, and tracking information. For these variables, there was normally a *true solution*. If it could not be reconstructed, these variables were at least made consistent with the data and the other identification variables. All inconsistencies within the identification or tracking variables were corrected.

Variables of the second type were those containing the responses of students, teachers, or principals. These may have included insoluble inconsistencies or impossible values; e.g., because respondents answered inconsistently. In these cases, for which there is no *true solution* (unless inconsistencies were identified as mispunches), decisions were made as to whether the answers should have been made consistent (in cases of filter and dependent questions), recoded to 'Invalid', or left as they were (but documented as an indication of the reliability of these questions). The final development of these rules depended very much on the real number of problems and rules applied in former IEA studies and the pilot and field trial phases of TIMSS.

6.3 IDENTIFICATION VARIABLES, TRACKING VARIABLES, AND INDICATORS

In most countries, nearly all problems related to identification variables could be solved by reviewing the tracking materials, inspecting the related files, or contacting the national centers. Only for a few cases in a few countries were a negligible number of problems corrected using the following rules.

For cases in which incorrect identification could be easily corrected, the identification variables were simply recoded consistently. If the hierarchical ID system was used correctly, it was possible in most cases to recreate IDs, either from other identification variables for the same observation or from linked observations. Observations that were added accidentally during data entry were deleted.

In the few cases in which identification variables (used for linkage between files) could not be recovered, they were replaced by 'Not administered' to guard against incorrect merging later. Whenever there were doubts about the linkage between the student questionnaire and the achievement items, the linkage was removed. However, it was possible to correct mismatches in most cases.

Inconsistently identified tracking variables (background information obtained from sampling forms that could not be verified from materials at the national center) were either:

- Derived from other observations (e.g., the date of testing or the stream), or
- Replaced by 'Invalid' when other sources were not available.

Indicator variables, i.e., Test Indicator Variables and Participation Indicator Variables, were made consistent with the data as long as the national center did not indicate that changes in the data were necessary. Lost or not-entered achievement booklets or questionnaires were identified and added to the files whenever possible.

6.3.1 SPLIT VARIABLES

In some test questions, respondents were allowed to check more than one option. In these cases, it was hard to differentiate between "Not checked" and "Missing." In all cases where at least one item was coded as "Checked" and no items were coded "Not checked," all missing items were recoded to "Not Checked." If all items in a list were coded as "Not checked," all variables were recoded to "Missing." This was possible because the item lists concerned were exhaustive. It was not possible to verify this with the instruments.

For other questions, respondents were asked to check "Yes" or "No" for several subquestions or "Zero" for items that were not applicable. In these cases, respondents may not have followed the directions on the questionnaire and may have answered whole question blocks only with "Yes" (or valid numbers when times or orders were asked) and "Missing." In these cases, missing data was recoded to "No" or "Zero." This affected a large number of cases and caused long reports, although it is not a serious problem and simply caused a standardization of the data.

6.3.2 FILTER AND DEPENDENT QUESTIONS

In cases where questions were explicitly designed as filter questions and corresponding dependent questions, the rule was that if a filter variable was answered negatively, then the dependent variables should have been coded "Not applicable." Two types of inconsistencies could occur. The first and more frequent case was that a respondent answered consistently, but the puncher assigned the "Missing" code to the "Not answered" questions instead of the correct "Not administered" code. These cases were replaced automatically.

In most cases where a respondent's answer to a dependent question was inconsistent with the filter question, the more precise dependent questions were given preference, because it is assumed that a respondent gave more thought to a question about how many hours per week he/she taught than just to a yes/no question. If these questions were answered consistently, the filter variable was recoded appropriately.

Exceptions to these rules arose when answers to the dependent questions made sense even though the filter questions indicated that the dependent questions should not be answered. For example, students were asked to answer questions on biology (*Student Questionnaire Population 2 (s) Q 32*) only if they were currently enrolled in biology courses. If they answered the questions focusing on the lessons, the corresponding filter was recoded.

If they answered only questions focusing on the subject but not on lessons or teaching, the evidence from the dependent questions was not strong enough to recode the filter, since they could know something about biology without being enrolled in a class. The formal inconsistency was then left in the data.

6.3.3 OTHER INCONSISTENCIES

The TIMSS Cleaning Program identified various other inconsistencies between variables or between observations. These were mostly logical inconsistencies between answers to questions that were not explicitly dependent.

Often questions were dependent but were not marked explicitly as dependent. For example, some questions were repeated in a different context (e.g., *Student Questionnaire Population 2 Q 22 & 25g*), or some answers to questions were possible only if other questions were answered in a certain way (e.g., *Student Questionnaire Population 2 Q25j - n*). In all cases where one answer was inconsistent with a majority of consistent implicitly dependent items, the single item was set to "Invalid." All cases where no recoding could be undertaken because of uncertainty were flagged.

In another typical case of inconsistency, the sum of numeric variables could have been obviously invalid, regardless of whether all variables contained valid values (e.g., if the number of girls enrolled in a school and boys enrolled in a school were both zero). These cases were set to "Invalid" or flagged if too many variables were involved and an unacceptable loss of information would have been caused.

All cleaning steps were documented so that all changes are reversible.

6.4 PRELIMINARY STATISTICS

The DPC prepared preliminary statistics consisting of univariate statistics on all questionnaires and cognitive items. The main objective of these statistics was to give the DPC, the International Study Center, and the national centers the opportunity to review the preliminary data and to check for possible errors or inconsistencies. The following statistics were produced:

- Item statistics and student scores
- Items statistics similar to the field trial item statistics and student scores to be merged to the final data sets
- Statistics on free-response items and reliability coding
- In addition to the item statistics, separate statistics on the free-response items, containing the item difficulty, the Rasch item difficulty, frequencies for all two-digit codes (the item statistics considered just the number of points obtained on the item), and statistics on the reliability of an item
- Univariate statistics

- A set of univariates for all background questionnaires, student and school data weighted with the sampling weights calculated by Statistics Canada.

6.4.1 INTERNATIONAL REVIEW OF DATA AFTER FIRST ANALYSES

After receiving the preliminary statistics, countries were asked to review their data and indicate necessary changes. Also during NRC meetings, country representatives were asked to respond to preliminary data and problems with the data.

The International Study Center and the DPC performed a parallel review of the data, not country by country as was done in the standard cleaning, but with international comparisons. The following checks were undertaken:

- Outliers were identified and corrected
- Variables with unexpectedly high values were identified and the corresponding instruments were reviewed
- Multiple choice questions for which one or more options were not used within a country were identified
- Typical inconsistencies that remained unchanged during standard cleaning were reviewed by comparing responses internationally, and possible solutions were identified
- Nationally defined codes for open-ended questions were recoded according to the international coding scheme
- Misprinted and mistranslated items as well as questions changed so as to preclude international comparison were deleted.

6.5 DATABASE CONSTRUCTION

The data files, prepared during data entry at the national center and sent to the DPC, should have been prepared according to the guidelines given in the international codebook provided with the DATAENTRYMANAGER software. The structure of these data files mirrored the structure of the tracking forms and instruments to facilitate data entry. To make the data files suitable for further analysis, the following data processing steps were taken:

- All Student Achievement Files were rearranged from the booklet-oriented structure (necessary for data entry) to a cluster-oriented structure. Redundant variables and variables necessary only for data entry were deleted so that the files became smaller.
- New codes were introduced for some variables (e.g., for the Participation Indicator Variables if the booklets were lost).
- Additional indicators were included (e.g., an overall Participation Indicator Variable).
- Information that could be derived from other sources was transcribed for variables corresponding to questions that were not administered (e.g., the information on teachers' personal backgrounds).

- School, classroom, and student weights were calculated and appended to the data file.
- Preliminary student scores were calculated for both math and science.

A detailed description of changes in the original files and newly introduced variables was provided with the national data documentation.

6.6 SUMMARY

Assembling, documenting, and standardizing the vast amount of data collected via the seven TIMSS tests and the multiple background questionnaires represents a daunting enterprise. Even though extreme care was taken in developing manuals and software for use by the 45 participating countries, the national centers, often inadvertently, introduced various types of inconsistencies in the data that needed to be thoroughly investigated. Thus, a series of steps was implemented to facilitate construction of the international database so that the data would be consistent across countries.

To ensure the availability of comparable, high-quality data for analysis, the data underwent an exhaustive cleaning process. That process involved several goals and procedures designed to identify, document, and correct deviations from the international instruments, file structures, and coding schemes. The process also emphasized consistency of information within national data sets and appropriate linking among the many data files.

The data cleaning process is an iterative one, involving double-checking and rechecking by the IEA Data Processing Center, the TIMSS International Study Center, and the national centers. The national centers were contacted regularly throughout the cleaning process and were given multiple opportunities to review the data for their countries. Considering the vast amount of available data, the process of database construction is an ongoing one for TIMSS and is expected to continue throughout the analysis process.

APPENDIX A

Figure A.1
Examples of Cross-Country Item Analysis

Population: 2 (B) Subject: 'Mathematics' Cluster: H Item: BSMMH07

Country	Correct Answer			Flags	Percentages for each alternative								Point biserials for each alternative								Rasch			Group Difficulties				International Mean	
	N	DIFF	DISCR		A	B	C	D	E	W	OMIT	NR	A	B	C	D	E	W	OMIT	NR	RDIFF	SE	FIT	MAL	FEM	LOW	UPP	IDIFF	IDISCR
AUS	5112	78.5	0.36	qG	3.3	9.8	78.5*	2.4	5.2		0.6	0.1	-0.21	-0.14	.36*	-0.10	-0.22		-0.07	-0.04	-1.14	0.04	1.07	77.9	79.1	70.3	78.7	68.1	0.38
AUT	2241	70.3	0.41	q	4.6	12.4	70.3*	3.4	6.6		2.2	0.1	-0.19	-0.16	.41*	-0.11	-0.19		-0.14	-0.06	-0.34	0.05	1.06	71.5	69.2	70.3	70.3	68.1	0.38
BFL	2123	82.9	0.37	qG	2.7	8.4	82.9*	2.5	2.9		0.5	0.0	-0.21	-0.19	.37*	-0.13	-0.17		-0.07	-0.01	-0.79	0.06	1.00	83.3	82.6	82.1	83.8	68.1	0.38
BFR	1841	72.5	0.39	q	8.0	12.2	72.5*	2.2	3.3		1.3	0.1	-0.24	-0.14	.39*	-0.09	-0.17		-0.11	-0.02	-0.60	0.06	1.05	73.6	71.7	70.1	74.6	68.1	0.38
BGR	1285	59.2	0.46	qG	8.2	16.2	59.2*	4.4	6.8		3.1	0.0	-0.17	-0.15	.46*	-0.12	-0.21		-0.19	0.00	0.06	0.06	1.04	59.2	59.2	55.5	62.9	68.1	0.38
CAN	6180	71.5	0.41	qG	5.5	10.9	71.5*	4.1	7.3		0.4	0.1	-0.21	-0.18	.41*	-0.11	-0.20		-0.08	-0.03	-0.83	0.03	1.00	72.0	71.2	69.3	73.7	68.1	0.38
CHE	3718	77.4	0.42	qG	4.2	9.7	77.4*	3.1	3.6		0.7	0.2	-0.25	-0.19	.42*	-0.12	-0.15		-0.11	-0.04	-0.67	0.04	1.02	77.8	77.1	78.4	82.6	68.1	0.38
CHE	588	79.8	0.38	qG	3.9	8.3	79.8*	2.2	4.3		0.5	0.0	-0.21	-0.17	.38*	-0.11	-0.19		-0.08	0.00	-0.73	0.11	0.98	80.9	78.5	78.9	80.6	68.1	0.38
CHE	570	75.8	0.36	qS	3.3	11.8	75.8*	3.2	3.9		0.4	0.4	-0.16	-0.15	.36*	-0.10	-0.20		-0.10	-0.10	-0.50	0.11	1.03	74.2	77.5	73.0	78.6	68.1	0.38
COL	1978	35.9	0.25		14.0	19.2	35.9*	7.7	14.8		5.8	1.6	0.00	-0.02	.25*	-0.05	-0.15		-0.15	-0.09	-0.58	0.05	1.07	36.3	35.7	35.4	36.5	68.1	0.38
CSK	2465	64.8	0.38	qsSG	5.5	19.7	64.8*	3.3	3.8		1.1	0.2	-0.21	-0.18	.38*	-0.10	-0.14		-0.08	-0.05	-0.02	0.05	1.11	67.3	62.3	62.7	66.9	68.1	0.38
CYP	2129	53.9	0.35		9.1	16.6	53.9*	6.0	10.1		2.5	0.0	-0.12	-0.10	.35*	-0.08	-0.19		-0.12	-0.03	-0.43	0.05	1.10	54.7	53.0	52.1	55.6	68.1	0.38
DEU	2258	70.2	0.41	q	5.3	9.8	70.2*	3.7	6.2		2.3	0.4	-0.22	-0.11	.41*	-0.09	-0.16		-0.19	-0.09	-0.86	0.05	1.03	70.4	70.7	66.8	73.6	68.1	0.38
DNK	2372	77.4	0.43	q	3.7	9.6	77.4*	3.1	3.7		1.4	0.4	-0.20	-0.20	.43*	-0.13	-0.20		-0.14	-0.04	-1.28	0.05	0.97	78.4	76.3	78.0	86.4	68.1	0.38
ESP	2813	54.7	0.40	qsS	8.7	19.0	54.7*	4.5	8.7		4.4	0.1	-0.19	-0.10	.40*	-0.09	-0.19		-0.15	0.00	-0.26	0.04	1.03	56.8	52.6	50.5	58.8	68.1	0.38
FRA	2245	75.6	0.35	qsS	6.1	11.4	75.6*	1.7	3.4		0.4	0.3	-0.20	-0.14	.35*	-0.06	-0.18		-0.05	-0.07	-0.90	0.05	1.04	78.1	73.2	70.7	80.7	68.1	0.38
GBR	1322	77.5	0.42	qG	2.6	11.6	77.5*	2.7	5.2		0.4	0.0	-0.20	-0.20	.42*	-0.12	-0.23		-0.10	0.00	-1.48	0.07	0.99	78.6	76.2	77.5	77.4	68.1	0.38
GRC	3017	45.4	0.42	G	17.8	14.5	45.4*	6.5	11.2		2.8	0.1	-0.17	-0.07	.42*	-0.07	-0.22		-0.12	0.00	0.02	0.04	1.06	46.2	44.5	38.1	52.6	68.1	0.38
HKG	2537	85.6	0.42	qs	1.7	5.5	85.6*	2.8	3.9		0.4	0.0	-0.23	-0.24	.42*	-0.09	-0.22		-0.05	0.00	-1.09	0.06	1.00	86.8	84.0	84.3	86.8	68.1	0.38
HUN	2224	61.5	0.46	qSG	9.2	15.5	61.5*	5.0	1.9		6.6	0.4	-0.24	-0.16	.46*	-0.18	-0.10		-0.17	-0.06	-0.06	0.05	1.02	62.9	60.4	55.0	68.3	68.1	0.38
IRL	2332	77.7	0.41	qS	3.7	11.2	77.7*	3.1	3.6		0.1	0.0	-0.23	-0.20	.41*	-0.11	-0.19		-0.06	-0.03	-1.07	0.06	1.01	77.7	77.8	75.3	80.3	68.1	0.38
IRN	2755	42.1	0.30	SG	17.5	20.3	42.1*	6.0	11.7		2.1	0.0	-0.13	-0.05	.30*	-0.04	-0.20		-0.01	0.00	-0.37	0.04	1.05	41.3	43.1	35.5	48.8	68.1	0.38
ISL	1388	73.1	0.40	q	3.6	13.4	73.1*	4.0	5.0		0.4	0.1	-0.18	-0.20	.40*	-0.15	-0.15		-0.06	-0.04	-1.37	0.07	0.98	73.5	72.7	69.3	77.5	68.1	0.38
ISR	518	63.3	0.38		7.9	16.6	63.3*	5.0	5.6		1.2	0.4	-0.16	-0.15	.38*	-0.18	-0.10		-0.10	-0.16	-0.18	0.10	1.10	65.0	64.0	63.3	63.3	68.1	0.38
JPN	3913	83.3	0.33	qSG	1.9	8.6	83.3*	2.4	3.7		0.0	0.0	-0.17	-0.14	.33*	-0.13	-0.22		0.00	0.00	-0.73	0.05	1.09	82.4	84.2	82.1	84.4	68.1	0.38
KOR	2160	91.3	0.45	qsG	1.7	3.1	91.3*	2.0	1.8		0.1	0.0	-0.19	-0.20	.45*	-0.21	-0.28		-0.04	0.00	-1.84	0.08	0.91	92.9	89.3	90.7	91.9	68.1	0.38
KWT	635	42.8	0.27		14.8	16.4	42.8*	7.1	14.5		4.1	0.3	-0.10	-0.03	.27*	-0.09	-0.14		-0.06	-0.10	-0.66	0.09	1.06	41.4	43.5			68.1	0.38
LTU	1882	45.2	0.39	sS	10.2	24.0	45.2*	5.6	6.9		6.5	0.4	-0.15	-0.09	.39*	-0.04	-0.18		-0.22	-0.06	0.02	0.05	1.06	42.3	47.8	39.7	50.7	68.1	0.38
LVA	1867	47.3	0.35	q	8.2	26.0	47.3*	4.6	9.4		3.9	0.6	-0.16	-0.06	.35*	-0.07	-0.21		-0.12	-0.05	0.06	0.05	1.10	46.7	47.9	43.5	51.2	68.1	0.38
MEX	4371	48.2	0.33	q	16.4	18.6	48.2*	4.0	10.8		1.2	0.4	-0.17	-0.06	.33*	-0.05	-0.19		-0.07	-0.04	-0.88	0.03	1.02	48.3	48.0	45.8	50.7	68.1	0.38
NLD	1546	78.4	0.36	q	3.2	13.8	78.4*	1.6	2.7		0.3	0.2	-0.27	-0.17	.36*	-0.08	-0.14		-0.11	-0.08	-0.98	0.07	1.06	78.4	78.2	74.9	82.0	68.1	0.38
NOR	2144	72.8	0.37	qG	4.2	12.3	72.8*	3.3	6.5		0.9	0.1	-0.21	-0.15	.37*	-0.08	-0.18		-0.12	-0.03	-1.12	0.05	1.02	72.0	73.3	70.2	74.8	68.1	0.38
NZL	2543	77.1	0.41	qsS	4.3	10.5	77.1*	3.1	4.7		0.1	0.0	-0.20	-0.20	.41*	-0.14	-0.19		-0.02	-0.04	-1.41	0.05	1.00	79.0	75.2	73.9	79.8	68.1	0.38
PHL	4478	49.3	0.15	qbBFSG	6.5	25.8	49.3*	4.7	13.2		0.4	0.1	-0.06	0.03	.15*	-0.07	-0.16		-0.03	-0.02	-0.92	0.03	1.20	50.2	48.7	49.7	49.0	68.1	0.38
PRT	2496	64.0	0.35	qsS	6.9	17.7	64.0*	3.0	6.1		2.2	0.2	-0.17	-0.13	.35*	-0.11	-0.15		-0.12	0.01	-1.14	0.05	1.01	67.5	60.4	60.9	67.0	68.1	0.38
ROM	2789	43.5	0.35	sS	16.6	16.9	43.5*	6.9	9.3		4.0	0.5	-0.12	-0.08	.35*	-0.07	-0.18		-0.12	-0.04	0.19	0.04	1.09	45.9	41.1	39.9	47.1	68.1	0.38
RUS	3036	54.4	0.31	qF	8.5	24.4	54.4*	2.6	5.6		3.3	0.4	-0.17	-0.04	.31*	-0.09	-0.20		-0.15	-0.03	0.20	0.04	1.19	54.9	53.9	51.5	57.3	68.1	0.38
SCO	2127	76.4	0.39	q	2.8	12.4	76.4*	2.9	4.8		0.3	0.0	-0.22	-0.17	.39*	-0.11	-0.23		-0.07	-0.02	-1.48	0.06	1.01	76.1	76.6	73.9	78.9	68.1	0.38
SGP	3096	85.5	0.36	qs	1.8	7.2	85.5*	2.2	3.2		0.1	0.0	-0.12	-0.24	.36*	-0.14	-0.17		-0.04	0.01	-0.56	0.05	1.02	84.5	86.4	82.0	88.2	68.1	0.38
SLV	2650	61.1	0.37	qFG	5.5	20.8	61.1*	4.0	5.8		1.4	0.2	-0.17	-0.16	.37*	-0.08	-0.18		-0.09	-0.05	0.03	0.05	1.13	60.3	61.8	58.6	63.6	68.1	0.38
SVN	2086	68.2	0.[illegible]	qs	5.5	17.3	68.2*	3.5	3.2		0.8	0.0	-0.21	-0.19	.38*	-0.09	-0.13		-0.06	0.00	-0.46	0.05	1.06	69.9	66.7	62.8	73.7	68.1	0.38
SWE	3277	73.9	0.38	q	3.3	13.5	73.9*	4.0	4.7		0.3	0.1	-0.20	-0.16	.38*	-0.13	-0.20		-0.08	-0.04	-0.96	0.04	1.05	73.4	74.4	74.6	80.0	68.1	0.38
USA	4108	69.7	0.37	qsS	5.4	14.5	69.7*	3.8	5.8		0.7	0.0	-0.21	-0.19	.37*	-0.11	-0.13		-0.07	0.01	-1.04	0.04	1.06	67.6	72.0	66.4	71.6	68.1	0.38

Figure A.2
Example of Graphical Displays of Cross-Country Item Statistics - Mathematics - Population 2

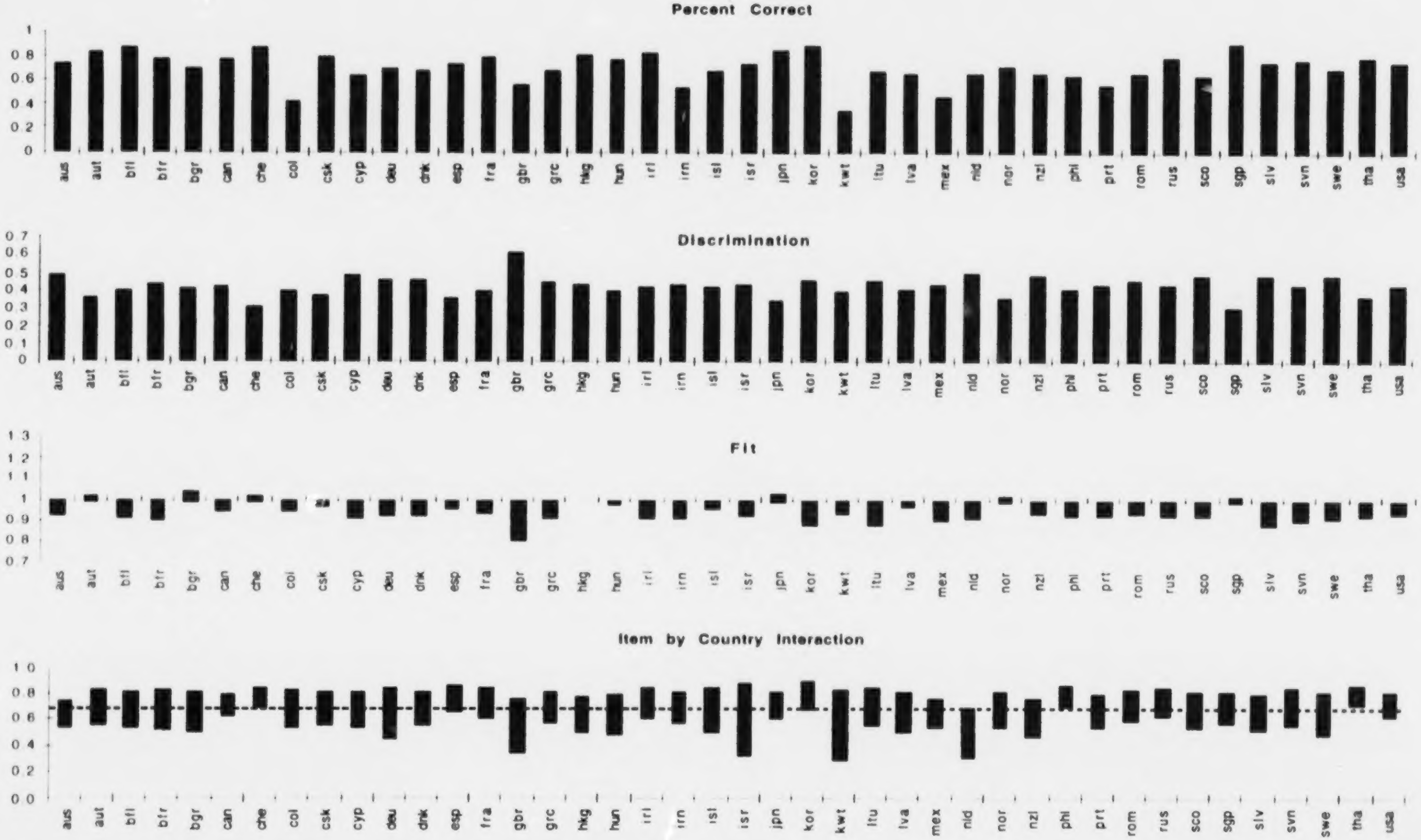


Figure A.3
Summary of Items with Poor Statistics for Some Countries

Country	Item by Country Interactions		Discrimination			Fit
	Easier than Expected	Harder than Expected	Non-key PB is Positive	Key PB is Negative	Ability not Ordered	Fit Large
	Tolerance=#Name					
tem: 119	*BSMSQ15*		*BSMS/WHICH IS NOT A CHEMICAL CHANGE (A)*			
DEU	☐	☐	☒	☐	☐	☐
HKG	☐	☐	☒	☐	☐	☐
ISL	☐	☐	☒	☐	☐	☐
ISR	☐	☐	☒	☐	☐	☐
NOR	☐	☐	☒	☐	☐	☐
PHL	☐	☐	☐	☐	☐	☐
tem: 120	*BSMSQ16*		*BSMS/HOW LONG TAKE LIGHT FROM STAR (D)*			
COL	☐	☐	☒	☐	☐	☐
CYP	☐	☐	☒	☐	☐	☐
DEU	☐	☐	☒	☐	☐	☐
GRC	☐	☐	☒	☐	☐	☐
HKG	☐	☐	☒	☐	☐	☐
ISR	☐	☐	☒	☐	☐	☐
KOR	☐	☐	☒	☐	☐	☐
MEX	☐	☐	☒	☐	☐	☐
ROM	☐	☐	☒	☐	☐	☐
THA	☐	☐	☒	☒	☐	☐

APPENDIX B

CHANGES MADE TO THE TIMSS INTERNATIONAL DATABASE FOR THE POPULATION 2 COGNITIVE ITEMS

ITEMS DELETED IN INDIVIDUAL COUNTRIES

All Countries

M09 (Staircase graph)

Austria

M05 Mathematics (Geometric Figure Half Turn)

Belgium (FR)

N05 Science (Cause Acid Rain)

Colombia

O12 (Composition of Earth)

Cyprus

F03 Science (Humans Interpret Senses)

H01 Science (Not Function Blood)

J01 Science (Describe Surface Earth Billions Years)

J07 Science (Warm-blooded Animals Differ Cold-Blooded Animals)

O11 Science (Which Chemical Change)

Q15 Science (Chemical Change)

Denmark

N16 Mathematics (Jan's Bag of Marbles)

France

N05 Science (Cause Acid Rain)

Greece

Q16 (Light from star)

Hungary

L17 Mathematics (Subtracting Fractions)

Q09 Mathematics (Add and Multiply Fractions)

Z01B & Z01C Science (Painting the Bridge)

Iceland

O04 Mathematics (Round Nearest Hundredth)

Q17 Science (Advantage Two Eyes)

Indonesia

A03 Mathematics (Actual length of box)

D10 Mathematics (Cost of printing greeting cards)

E01 Mathematics (Graph distance and time of hike)

J14 Mathematics (Divide with decimals)

J17 Mathematics (Map of Oxford and Smithville)

M04 Mathematics (Which number largest?)

N11 Mathematics (Rounding number of trees planted in park)

N17 Mathematics (Fuel in tank of car)

P17 Mathematics (Table of temperatures)

A12 Science (Fast moving river)

I11 Science (Features of all insects)

J01 Science (Surface of earth over billions of years)

M11 Science (Food web)

W01A Science (Good place farming)

Iran

F04 Science (Wash Away Soil)

N09 Science (Filtration)

Israel

B05 (Which feature is located) dropped for booklet 1 only.

Japan

J11 Math (Angles in parallelogram)

N17 Math (Car consumes fuel)

U02A and U02B (Drawing in grid)

Kuwait

D10 Mathematics (Cost Greeting Cards)

F07 Mathematics (Average Speed Runner)

I01 Mathematics (Whole Numbers Add to 81)

J05 Science (Solar Radiation Sunburn)

J18 Mathematics (Table x and y)

O04 Mathematics (Round Nearest Hundredth)

Latvia

E12 Science (Caves in Limestone)

Lithuania

F11 Mathematics (4 Times Number 48, 1/3 Number)

Mexico

J1 Science (Describe Surface Earth Billions Years)

Philippines

I03 Mathematics (Number 750mL Bottles from 600 L Water)

J02 Science(Species on Earth Shortest Time)

Romania

A03 Mathematics (Length Box Nearest Centimeter)

Singapore

U02A and U02B (Drawig in grid)

Slovak Republic

K10 Science (How Does Air Exist)

Thailand

K04 Mathematics (X/2 <7)

P04 Science (When animal hibernates)

Q16 Science (Light Nearest Star)

United States

I13 Science (Best Thermometer Body Temperature)

OPTIONS RELABELED IN ITEMS

Greece

N08 Science (Girls Balance on Seesaw): Options A and B are switched and options C and D are switched

Indonesia

B10 Mathematics (Which smallest number): Options B and D should be switched

M02 Mathematics (Lines of symmetry for rectangle): Options A and C should be switched

Korea

B02 Science (Chemical Energy Released Car): Options B and C are switched

C02 Mathematics (Graph Distribution of Crops): Options C and D are switched

P12 Mathematics (Mark's Garden): Options B and C are switched, Options A and D are switched

Latvia

C01 Mathematics (Stack Blocks Different Volume): Options A and B are switched

K01 Mathematics (Same Fraction Shaded): Options C and D are switched

M02 Mathematics (Diagram Lines of Symmetry Rectangle): Options A and C are switched and options B and D are switched

FREE-RESPONSE ITEM RECODING

Science

K10 (How Air Exists)

Categories 70 and 71 should both = 70. Training team found it difficult to distinguish between these categories.

L04 (Two Machines) -- Only 20s have positive PB, so recode:

20 = 10, 21 = 11, 29 = 19

10 = 74, 11 = 75, 12 = 76, and 19 = 79

M11 (Food Web) -- Only 30s have positive PB, so recode:

10 to 13 = 71

20 to 25 = 72

30 = 10 and 31 = 11

Y01 (Energy in a Lamp) -- Only 20s have positive PB, so recode:

20 = 10, 21 = 11, 22 = 12, 29 = 19

10 = 73, 11 = 74, 19 = 75

Y02 (Warming Snowballs) -- Typo in category 21 in coding guide, so recode:

21 = 19

J03 (Molecules)

19=10

M12 (Use Ammemeter)

19=10

O14 (Sun and Moon)

20 -> 10 ; 29 -> 19 ; 10 -> 72 ; 11 -> 73 ; 19 -> 74 due to e-mail from Ina

Q18 (Melting Ice)

19=10 and 29=20

Mathematics

L16 (Solve for X)

19=10

M06 (Ratio Boys/Girls)

19=10

M08 (Decimal Multip.)

19=10

Q10 (Degrees of Angle)

19=10

R13 (Spent Fraction Total Money)

Recode Category 74 = 79. Category 74 has error in guide (28 instead of 280). This does leave gap in response range 70 to 73, 75, 79.

S01A (Congruence...)

19=10

S02A (Area of Figure)

19=10

T01A (Apples in Box)

29=20

T02A (Pattern boxes)

19=10

U01A (Estimate time)

19=10

U02A (Draw Rectangle)

19=10 and 29=20

U02B (Draw Rectangle)

19=10 and 29=20

ITEMS THAT COUNTRIES HAVE DELETED OR OMITTED

France

V01 Mathematics : weight of Dolphin (had translation problem)

Russia

I19 Science -- Table Oxygen in Pond

APPENDIX C

Form 1 **TIMSS Participation**

See Section 5.5 of Sampling Manual.

TIMSS Participant: ____________________

National Project Coordinator: ____________________

1 Specify the populations that will be investigated in your country. For each population, specify the usual start date of the school year, and the expected date of testing for the main survey.

	Usual start date of the school year:	Expected date of testing for the main survey:	Usual end date of the school year:
○ Population 1			
○ Population 2			
○ Population 3			

2. Describe the age and birthdate rules for entering primary school in your country.

3. Describe the grade structure of early primary school in your country *(for example, nursery, kindergarten, grades 1 through 6, etc.)*.

Use additional sheets if necessary

Population 1

Describing National Desired Target Population

Form 2/part 1

See Section 6.1 of Sampling Manual.

TIMSS Participant: ______________________________

National Project Coordinator: ______________________________

> *The international desired target population includes all students enroled on a full-time basis in the two adjacent grades that contain the largest proportion of students in the age 9 cohort at the time of testing.*

1a. Do you plan to sample two classrooms per grade? ○ yes *or* ○ no

1b. Do you plan to subsample within classrooms? ○ yes *or* ○ no

2a. Specify the adjacent grades selected for the international desired target population:
○ grades 3 and 4 ○ grades 2 and 3 ○ other *(specify and explain):*

____

2b. Specify the percent coverage of the age cohort in the grade pair: [%]

2c. Total national enrolment in the grades specified in 2a: [a]

3a. Describe the coverage of the national desired target population emphasizing any differences from the international desired target population:

____

____

Total enrolment in the desired target population: [b]

3b. Describe the population(s) to be excluded from the national desired target population *(if applicable):*

____

____

Total enrolment excluded from national desired target population *(box [a] - box [b]):* [c]

Percentage *(box [c] ÷ box [a]):* [d] %

4. Describe the data source: ______________________________

____

Use additional sheets if necessary

Population 2

Form 2/part 1 Describing National Desired Target Population

See Section 6.1 of Sampling Manual.

TIMSS Participant: ____________________

National Project Coordinator: ____________________

> *The international desired target population includes all students enroled on a full-time basis in the two adjacent grades that contain the largest proportion of students in the age 13 cohort at the time of testing.*

1a. Do you plan to sample two classrooms per grade? ○ yes *or* ○ no

1b. Do you plan to subsample within classrooms? ○ yes *or* ○ no

2a. Specify the adjacent grades selected for the international desired target population:
○ grades 7 and 8 ○ grades 6 and 7 ○ other *(specify and explain):*

____

2b. Specify the percent coverage of the age cohort in the grade pair: [%]

2c. Total national enrolment in the grades specified in 2a: [a]

3a. Describe the coverage of the national desired target population emphasizing any differences from the international desired target population:

____

____

Total enrolment in the desired target population: [b]

3b. Describe the population(s) to be excluded from the national desired target population *(if applicable):*

____

____

Total enrolment excluded from national desired target population *(box [a] - box [b]):* [c]

Percentage *(box [c] ÷ box [a]):* [d %]

4. Describe the data source: ____________________

____

Use additional sheets if necessary

Form 2/part 1

Population 3

Describing National Desired Target Population

See Section 6.1 of Sampling Manual.

TIMSS Participant: ____________________

National Project Coordinator: ____________________

The international desired target population includes all students who are enroled on a full-time basis in one of the final segments of public or private secondary education.

OO = Math generalists and physics generalists
MO = Math specialists and physics generalists
OP = Math generalists and physics specialists
MP = Math specialists and physics specialists

1. Identify and describe the Population 3 subsystems in your country:

130

Use additional sheets if necessary

2. Retention rates are important in describing differences between school systems. Please tell us the percentage of each age cohort enroled in school in your country. (This is an update of information provided in the Participation Survey).

9 __________
13 __________
14 __________
15 __________
16 __________
17 __________
18 __________
19+ __________

3. Define the mathematics and/or physics specialists under investigation:

Math (M)

Physics (P)

Use additional sheets if necessary

4. Describe the distribution of student groups by sub-system, and type of school (1 form per sub-system).

Population Sub-System:		
Student Groups	% of Enrolment in Sub-system	Type of School(s)
OO		
MO		
OP		
MP		

Population Sub-System:		
Student Groups	% of Enrolment in Sub-system	Type of School(s)
OO		
MO		
OP		
MP		

Use additional sheets if necessary

Form 2/part 2 National Defined Target Population

To be completed for EACH population listed in Form 1. See Section 6.2 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1):*

1. Number of students in national desired target population *(from box [b] on Form 2/part 1):* [a]

2. School exclusions:

Reason for exclusion:	No. of students:
TOTAL	[b]

Percentage of exclusions *(box [b] ÷ box [a]):* [c]

%

3. Number of students in national defined target population *(box [a] - box [b]):* [d]

4. Within-school exclusions:

Reason for Exclusion:	Expected No. of students:
TOTAL	[e]

5. Describe the data source:

Use additional sheets if necessary

Form 3 Stratification Variables

To be completed for EACH population listed on Form 1. See Section 7 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1):*

Part 1

List and describe the variables used for design domains:

No.	Description:	Number of levels:

Part 2

Provide details about enrolment in single grade and multi-grade schools *(for Populations 1 and 2 only):*

	Enrolment in lower grade:	Enrolment in upper grade:	Total Enrolment:
Schools with both grades:			
Schools with only upper grade:	———		
Schools with only lower grade:		———	
TOTAL			

Part 3

List and describe the variables used for implicit strata:

No.	Description:	Number of levels:

Use additional sheets if necessary

Form 4/part 1 **Stratification Variables (cont.)**

To be completed for EACH population listed on Form 1. See Section 8 of Sampling Manual.

TIMSS Participant: ____________________

National Project Coordinator: ____________________

Assessment population *(from Form 1)*: ____________________

1. Stratum Number and name: ☐
2. Enter the minimum cluster size (mcs) to be used: ☐
3. Specify the source of information used to determine the coefficient of intraclass correlation (rho):

   ____

   ____

   Enter rho: ☐
4. Specify the precision requirements *(i.e. 95% confidence limits of ±0.2s for estimated means)*:

   ____

   ____

---

1. Stratum Number and name: ☐
2. Enter the minimum cluster size (mcs) to be used: ☐
3. Specify the source of information used to determine the coefficient of intraclass correlation (rho):

   ____

   ____

   Enter rho: ☐
4. Specify the precision requirements *(i.e. 95% confidence limits of ±0.2s for estimated means)*:

   ____

   ____

Use additional sheets if necessary

Form 4/part 2 Type of Sample Frame

To be completed for EACH population listed on Form 1. See Section 9.4 of Sampling Manual.

TIMSS Participant: __________

National Project Coordinator: __________

Assessment population *(from Form 1)*: __________

1. Describe the type of sampling frame to be used:

○ Single level *(check one)*:
Type A ○
Type B ○
Type C ○

○ Double level *(check one)*:
Type D ○
Type E ○

○ Other *(specify)*: __________

2. Describe the measure of size to be used *(i.e. total enrolment in target grades for 1990-1991 school year)*:

____

3. If double-level frame is to be used provide preliminary description of information available to construct this frame. The ICC will provide support, if necessary, to assist the NPC in the construction and use of a double-level frame.

____

Use additional sheets if necessary

Form 5/part 1 Schools Excluded from the Sampling Frame

To be completed for EACH population listed on Form 1. See Section 9.5 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1)*:

Identification number of excluded school	REASON FOR EXCLUSION	Measure of size of excluded school
TOTAL		

Use additional sheets if necessary

Form 5/part 2 Recording the Formation of Pseudo-schools

To be completed for EACH population listed on Form 1. See Section 9.5 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1)*:

Identification Number of Superschool*	Identification Numbers of Collapsed-schools Forming the Pseudo-school					Measure of Size of Pseudo-school
	School #1	School #2	School #3	School #4	School #5	
TOTAL						

* Retain the identification number of the largest school making up the pseudo-school.

Use additional sheets if necessary

Form 6/part 1 Strata for Defined Target Population

To be completed for EACH population listed on Form 1. See Section 9.8 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1)*:

(1) Design Domain (D.D.)	(2) Stratum (Name)	(3) Population of Schools (Number)	Population of Students	
			(4) Number	(5) Percentage of: Pop. D.D.

Use additional sheets if necessary

Form 6/part 2 **Strata for Defined Target Population**

To be completed for EACH population listed on Form 1. See Section 9.8 of Sampling Manual.

TIMSS Participant:

National Project Coordinator:

Assessment population *(from Form 1)*:

(1) Design Domain	Stratum		Sample of Schools	Expected Sample of Students	
	(2) Name:	(3) mcs	(4) Number:	(5) Lower Grade	(6) Upper Grade

Use additional sheets if necessary

Form 7 Selecting the Sample of Schools

TIMSS Participant:

Assessment population *(from Form 1 — including Design Domain and Explicit Stratum) :*

(1) School List Number	(2) Pseudo-school (✓)	(3) Measure of Size (from sorted sampling frame)	(4) Cumulative Measure of Size	(5) Sampled (✓)
Total from previous page.				

Use additional sheets if necessary

Form 8 Identifying the Sample of Schools

To be completed for EACH population listed on Form 1. See Section 10.3 of Sampling Manual.

TIMSS Participant:

Assessment population *(from Form 1 — including Design Domain and Explicit Stratum):*

S [a] Total Measure of Size	D [b] Desired Sample	I [c] Sampling Interval	RN [d] Random Number

(1) Line Number	(2) Selection Numbers

Use additional sheets if necessary

Form 9

Populations 1 and 2
School Tracking Form

Use one form for each school from the designed sample along with its replacement schools. See Section 10 of Sampling Manual.

Population: ______________________________ TIMSS Participant: ______________

[a] Total Measure of Size

	(1) School ID	(2) Name, Address and Phone Number of School	(3) Name and Phone Number of School Coordinator	(4) Measure of Size	(5) Status*	(6) Date Materials Sent Date Materials Returned	(7) Date of Assessment
†1.							
†2.							
†3.							

Use additional sheets if necessary

†Number 1 is the school from the designed sample. Number 2 is the first replacement school. Number 3 is the second replacement school.
*Enter "N" for non-participating. A check-mark (√) indicates participation.

Populations 1 and 2
Class Tracking Form

Population: ☐ 1 ☐ 2 TIMSS Participant: ____________

To be completed for participating schools listed on the Form 9. See Section 11 of Sampling Manual.

[a] School Name & Address: ____________

____

[b] School Coordinator & Contact: ____________

[c] School ID	[d] Minimum Cluster Size	[e] Target Grade	[f] Random Start	[g] Sampling Interval	[h1] First Selected Classroom	[h2] Second Selected Classroom

(1) Line Number	(2) Class ID	(3) Stream or Track	(4) Location (Room)	(5) Total Number of Students	(6) Number of Excluded Students	(7) Number of Students Eligible for Testing	(8) Cumulative Count	(9) Sample

Use additional sheets if necessary

APPENDIX D

TIMSS Quality Control Monitors

Argentina
Ana Lia Quiroz

Australia
Martin Caust

Austria
Gudrun Queitsch

Belgium (Flemish)
Norbert Delagrange

Belgium (French)
Simeon Simenya

Bulgaria
Petia Assenova

Canada
Robert J. Wilson
Tammy Conacher

Colombia
Jairo Alvarez

Cyprus
Christos Theophilides

Czech Republic
Helena Stehlikova

England
David Harris
Derek Foxman

France
Jean Geoffroy

Latvia
Andris Grinfelds

Lithuania
Pranas Gudynas

Mexico
Margarita L. Gutierrez-Talamas

Netherlands
Annebert Lammerts

New Zealand
Ian Livingstone

Norway
Astrid Eggen Knutsen

Greece
Philippos Vlachos

Hong Kong
M.C. Hung

Hungary
Judit Rosza

Iceland
Fridrik H. Jonsson

Indonesia
Djemari Mardapi

Iran
Mohammed Jafar Javadi

Israel
Ruth Raz

Italy
Silvia Guigni

Japan
Hisashi Kawai

Philippines
Filma G. Brawner

Portugal
Maria Jose Pagarete Cordeiro

Romania
Mihaela Muresan

Russian Federation
Eugene K. Straut

Scotland
Donald Gray

Slovak Republic
Juraj Vantuch

Slovenia
Petar Pavesic

South Africa
Fred Shaw

Spain
Blanca E. Valtierra Arrizabalaga

Sweden
Rune Palsson

Switzerland
Christian Langenegger

Thailand
Somchai Chinatrakool

Ukraine
Victor N. Akhmetov

United States
George Hall
Irving Broudy

TIMSS QUALITY CONTROL MONITOR TRAINING

Date	Location	Country	QCM
February 13 - 15, 1995	**Slough, England**	Austria	Gudrun Queitsch
		England	David Harris Derek Foxman
		Hungary	Judit Rosza
		Iceland	Fridrik H. Jonsson
		Philippines	Filma G. Brawner
		Scotland	Donald Gray
		Slovenia	Petar Pavesic
		Sweden	Rune Palsson
March 6 -7, 1995	**Enschede, The Netherlands**	Belgium (Fl)	Norbert Delagrange
		Belgium (Fr)	Simeon Simenya
		Cyprus	Christos Theophilides
		Greece	Philippose Vlachos
		Lithuania	Pranas Gudynas
		Netherlands	Annebert Lammerts
		Norway	Astrid Eggen Knutsen
		Thailand	Somchai Chinatrakool
		United States	George Hall
April 3 - 4, 1995	**Paris, France**	Bulgaria	Petia Assenova
		Canada	Robert J. Wilson Tammy Conacher
		Colombia	Jairo Alvarez
		Czech Republic	Helena Stehlikova
		France	Jean Geoffroy
		Hong Kong	M.C. Hung
		Indonesia	Djemari Mardapi
		Israel	Ruth Raz
		Italy	Silvia Guigni
		Japan	Hisashi Kawai
		Latvia	Andris Grinfelds
		Mexico	Margarita L. Gutierrez Talamas
		Portugal	Maria Jose Pagarete Cordeiro
		Romania	Mihaela Muresan
		Russian Fed.	Eugene K. Straut
July 27, 1995	**Wellington, New Zealand**	Australia	Martin Caust
		New Zealand	Ian Livingstone
September 13, 1995	**Philadelphia, PA**	Argentina	Ana Lia Quiroz

APPENDIX E

Results of the Quality Assurance Monitors' Interviews with the National Research Coordinators

Each quality control monitor visited the TIMSS national center in their country to interview the National Research Coordinator (NRC) about aspects of their data collection activities. The interview data which follows is based upon interviews conducted with 43 NRCs.

A. Sampling

A.1. Were you able to select a sample of schools and students within the schools using the Survey Operations Manual and Sampling Manual provided by the TIMSS Study Center?

		Number of NRCs			
		Yes	No	Not Applicable	No Response
a.	Population 1	24	1	16	2
b.	Population 2	35	5	1	2
c.	Population 3: Generalist	19	4	16	4
d.	Population 3: Math Specialist	17	2	19	5
e.	Population 3: Physics Specialist	18	2	18	5

A.2. If the answer to any of the above is no, please ask the NRC to explain.

Eight NRCs provided explanations for not using the procedures in the Survey Operations Manual and Sampling Manual.

		Number of Comments
a.	Selection was performed by person/group other than NRC (external authority)	2
b.	National circumstances necessitated a change in sampling procedures	2
c.	All schools were included in the sample	2
d.	Accurate class lists were not available.	2

A.3. Did you use the Sampling and Operations software provided by the TIMSS Study Center to select classes or students?

Number of NRCs		
Yes	No	No Response
17	25	1

A.4. If yes, was it helpful?

Thirteen of the 17 NRCs who used the software said the Sampling and Operations software provided by the TIMSS Study Center was helpful. Of the 17 NRCs who used the software, 9 provided comments. These comments are tabulated below.

		Number of Comments
a.	Some problems with software were noted	7
b.	Software systematized the procedure	1
c.	Only for sampling students and teachers and not for sampling classes	1
d.	Not much because sampling procedures for these countries were quite simple	1

A.5. If no, why not?

Twenty four NRCs provided explanations for not using the TIMSS software. These are tabulated below. Note that some NRCs provided more than one comment. Five NRCs reported that problems with the TIMSS software resulted in their selection or development of alternatives.

		Number of Comments
a.	Used other software already in place	10
b.	Software provided by TIMSS presented problems	12
c.	Sampling was conducted manually	3
d.	Sampling out of NRC control	2

A.6. Were there any conditions or organizational constraints that necessitated deviations from the basic TIMSS sampling design?

		Number of Comments			
		Yes	No	Not Applicable	No Response
a.	Population 1	3	23	15	2
b.	Population 2	10	31	-	2
c.	Population 3: Generalist	8	14	17	4
d.	Population 3: Math Specialist	4	14	20	5
e.	Population 3: Physics Specialist	6	4	18	5

A.7. If the answer to any of the above was yes, please ask the NRC to explain.

Sixteen NRCs gave explanations. These explanations mostly referred to Population 2 and Population 3. The explanations tended to highlight national, regional or school level features. The reasons given were diverse, reflecting the uniqueness of each education system.

		Number of Comments
a.	National education structure necessitated change	9
b.	School level organization such as student groups or curriculum structures necessitated change	5
c.	Sampling was conducted to avoid clashes with existing studies or programs	1
d.	Sampling was conducted in accordance to directions given by external authority	1

A.8. In terms of the complexity of the procedures and number of personnel needed, how would you describe the process of sample selection?

A total of eight NRCs identified at least one population as being very difficult to sample.

		Number of NRCs		
		Population 1	Population 2	Population 3
a.	very difficult	2	5	3
b.	somewhat difficult	10	17	11
c.	not difficult at all	14	19	12
d.	not applicable	-	2	-
		n =26	n = 43	n =26

A.9. If very difficult, please ask the NRC to explain.

Fifteen NRCs made comments on the process of sample selection. This included 7 NRCs who had not identified the sampling process as being very difficult. One NRC who had responded that the process of sample selection was "not difficult at all" mentioned that the procedures were not difficult, but there were major difficulties with the practical application of the procedures. The explanations given included the complexity and size of the country's education system, inadequate resources and the lack of student lists.

		Number of Comments
a.	Difficulties in obtaining sample (lack of class lists, need to sample a sub-population)	8
b.	Computer difficulties related to either software or hardware	4
c.	Inadequate resources for the study (either staffing or financial)	4
d.	Communication problems between NCR and regional or school system including teachers and principals unwilling to cooperate, curriculum pressures	3

B. Working With the School Coordinators

B.1. Have all the School Coordinators for your sample been contacted?

Number of NRCs		
Yes	No	No Response
38	4	1

B.2. If yes, have you sent them materials about the testing procedures?

Number of NRCs		
Yes	No	No Response
31	11	1

B.3. If the answer to B.1 and or B.2 is no, please ask the NRC to explain when, how, and where the School Coordinators will be contacted and how they will learn about their responsibilities.

Fourteen NRCs offered comments.

		Number of Comments
a.	Materials will be sent or are in the process of being sent	7
b.	Testing is being conducted by another agency (for example, students at a university, regional staff, head teachers)	6
c.	School coordinators are members of the national center staff	2
d.	NRC staff to contact school coordinators by telephone	2

B.4. Did you have formal training sessions for the School Coordinators?

Number of NRCs		
Yes	No	No Response
21	22	-

With the exception of a single no-response, all NRCs reported making changes to the TIMSS documents. Of these, eight NRCs made only cultural adaptations, the international versions of the instruments were prepared in English.

C. Translating the Documents

C.1. How difficult was it to translate and/or adapt the test booklets?

		Number of NRCs
a.	very difficult	3
b.	somewhat difficult	25
c.	not difficult at all	14
d.	no response	1
		n = 43

C.2. Did you go use your own staff or outside experts to translate the test booklets?

		Number of NRCs
a.	used own staff	10
b.	used outside experts	7
c.	used a combination	24
d.	N/A (no translation or adaptation)	1
e.	no response	1

C.3. Did you go through the process of submitting your test booklets and receiving a Translation Verification Report from the Study Center?

Two of the eight NRCs whose education system used English did not select a response category.

Number of NRCs		
Yes	No	No Response
33	8	2

C.4. If no, please ask the NRC to explain.

Eight NRCs gave explanations for not submitting a Translation Verification Report. In each case the test had to be translated and adapted to the local school system. The usual reason for not submitting the Translation Verification Report was lack of time.

		Number of Comments
a.	Time constraints	6
b.	Process done by a different agency	1
c.	Booklets were sent to printers prior to receipt of TVR	1
d.	Yet to receive TVR	1

C.5. How difficult was it to adapt the questionnaires?

		Number of NRCs
a.	very difficult	10
b.	somewhat difficult	23
c.	not difficult at all	10
		n = 43

C.6. If very difficult, please ask the NRC to explain.

Fourteen NRCs offered comments explaining their difficulty in translating and adapting the questionnaires. Four NRCs who indicated that the process was "somewhat difficult" also offered comments.

The main difficulty NRCs reported centered upon the disparity between the educational context of the country and that assumed by TIMSS. The difficulties tended to relate to school- and teacher-level features rather than to the student questionnaires. For example, "the student questionnaires are not difficult at all to adapt, but teacher and school questionnaires are very difficult."

		Number of Comments
a.	Questions did not match country's educational system	8
b.	Questions were unclear	2
c.	Questionnaires were too long	1
d.	Questionnaires were badly designed	1
e.	Questions were not applicable	1
f.	Terminology caused problems	1

C.7. How difficult was it to adapt the *Test Administrator Manual*?

		Number of NRCs
a.	very difficult	3
b.	somewhat difficult	11
c.	not difficult at all	26
d.	no response	3
		n= 43

C.8. If very difficult, please ask the NRC to explain.

Although only 3 NRCs indicated that it was very difficult to adapt the Test Administrator Manual, an additional 11 NRCs offered comments upon the adaptation process. Most NRCs who commented upon the content and length of the Test Administrator Manual stated that a simplified, more concise version was developed for their particular context.

		Number of Comments
a.	Manual was too long	8
b.	Manual was overloaded, too detailed, manual was simplified	8
c.	Combined the TA and SC manual	2

C.9. How difficult was it to adapt the *School Coordinator Manual*?

		Number of NRCs
a.	very difficult	4
b.	somewhat difficult	8
c.	not difficult at all	24
d.	no response	7
		n = 43

C.10. If very difficult, please ask the NRC to explain.

Three of the four NRCs who responded that the School Coordinator Manual was difficult to adapt offered explanations. In addition, two other NRCs commented upon the adaptation task.

		Number of Comments
a.	Manual was inappropriate, had to abbreviated	2
b.	The TA and SC manuals were combined	2
c.	Manual was not adaptable to this country's school system	1

C.11. Did you translate or do you plan to translate the *Coding Guides for Free Response Items*?

Number of NRCs		
Yes	No	No Response
19	24	-

D. Assembling and Printing the Test Materials

D.1. Were you able to assemble the test booklets according to the instructions provided by the Study Center?

		Number of NRCs		
		Yes	No	No Response
a.	Population 1	24	2	4
b.	Population 2	40	1	2
c.	Population 3	24	-	8

D.2. If no, please ask the NRC to explain.

Three NRCs gave explanations for not assembling the test booklet according to the instructions provided by the Study Center. These explanations were:

(a) For Population 1, the item numbering system was changed to avoid confusion. (Comment made twice).

(b) Instead of one booklet per pupil at Population 3, two separate test booklets were constructed, one for before the break and one after the break.

D.3. How difficult was it to assemble the test booklets?

		Number of NRCs
a.	very difficult	2
b.	somewhat difficult	11
c.	not difficult at all	30
		n = 43

D.4. If very difficult, please ask the NRC to explain.

In addition to the two NRCs who identified the test booklet assembly as "very difficult", five other NRCs offered comments. The NRCs who reported the task to be "very difficult" commented that it was very time consuming, they had insufficient personnel, and the graphics and labels were difficult to lay out.

		Number of Comments
a.	Too little time	5
b.	Lack of personnel	3
c.	Graphics and labeling were difficult	1

D.5. Did you conduct the quality assurance procedures for checking the test booklets during the printing process?

Number of NRCs		
Yes	No	No Response
35	7	1

D.6. If no, please ask the NRC to explain.

Eight NRCs offered explanations for not conducting quality assurance procedures.

		Number of Comments
a.	Test booklets checked by the printers	2
b.	Printing process not yet completed	2
c.	Not checked due to shortage of time	1
d.	NRC trusted the quality of the printers	1
e.	Shortage of staff due to budget limitations	1
f.	This will be done before packing the materials	1

D.7. Were any errors detected during the printing process?

Number of NRCs		
Yes	No	No Response
19	22	2

D.8. If yes, what was the nature of the error?

		Number of NRCs		
		Yes	No	No Response
a.	print quality	12	5	2
b.	pages missing	8	9	2
c.	page order	6	10	3
d.	upside down pages	-	15	4

D.9. Did you follow procedures to protect the security of the tests during the assembly and printing process?

Number of NRCs		
Yes	No	No Response
38	4	1

D.10. If no, please ask the NRC to explain.

Four NRCs gave explanations.

		Number of Comments
a.	NRC did not feel the need for special security measures	1
b.	Booklets were printed by external printers	1
c.	Booklets were printed internally	1
d.	The procedures were deemed too expensive to follow	1
		Total = 4

D.11. Did you discover any potential breaches of security?

Number of NRCs		
Yes	No	No Response
1	42	-

D.12. If yes, please explain and include whatever steps were taken to remedy the problem.

There was no comment offered by the NRC who reported a potential break of security.

D.13. Did you print the testing materials in-house or did you use an external printer? (check one for each)

		Number of NRCs		
		In-House	External	In-House & External
a.	Test Booklets	10	29	4
b.	Questionnaires	12	24	7
c.	Manuals (TA, SC, Coding)	26	12	5

E. Packing, Shipping, and Returning the Testing Materials

E.1. On what date did you or do you plan to begin testing?

	Number of NRCs
No date specified	2
February, 1995	8
March, 1995	10
April, 1995	7
May, 1995	13
July, 1995	2
October, 1995	1

The modal date was May, 1995

E.2. In packaging the assessment materials for shipment to schools, did you detect any errors in any of the following items?

		Number of NRCs			
		No Errors or Not Used	Errors Found Before Dist.	Errors Found After Dist.	No Response
a.	Supply of test books	30	2	1	8
b.	Supply of student questionnaires	27	4	2	8
c.	Student Tracking Forms	30	2	1	8
d.	Teacher Tracking Forms	28	2	-	11
e.	Student-Teacher Linkage Forms	27	-	-	14
f.	Test Administrator Manual	30	1	-	10
g.	School Coordinator Manual	27	2	-	12
h.	Supply of Teacher ..Questionnaires	26	3	1	11
i	School Questionnaire	31	2	1	7
j.	Test book ID labels	30	2	-	9
k.	Sequencing of books or questionnaires	30	1	2	8
l.	Return labels	29	-	1	11
m.	Self-addressed post-cards for test dates	27	-	-	14

E.3. Did concerns about confidentiality restrict your freedom to put student names on the booklet covers?

Number of NRCs		
Yes	No	No Response
18	20	3

E.4. Do you plan to or have you already established a procedure requiring schools to confirm receipt of the testing materials and verification of the contents?

Number of NRCs		
Yes	No	No Response
21	12	8

E.5. What date have you specified as the deadline for the return of materials from the schools?

	Number of NRCs
No date specified	14
March, 1995	7
April, 1995	1
May, 1995	8
June, 1995	11
July, 1995	1
December, 1995	1

The modal date was June, 1995

F. Coding Free-Response Questions

F.1. Who will primarily be coding your free-response questions?

		Number of NRCs
a.	own staff	7
b.	teachers	7
c.	university students	5
d.	combination of the above	24
		n = 43

F.2. How many coders do you plan to use for the coding of the free-response questions?

	Number of NRCs
1 to 10	17
11 to 20	14
21 to 30	3
31 to 40	4
41 to 50	4
51 or more	1
No Response	3

F.3. Have you selected your coders for the free-response questions?

Number of NRCs		
Yes	No	No Response
32	10	1

F.4. If yes, have you trained the coders?

Number of NRCs		
Yes	No	No Response
19	15	8

F.5. Have you scheduled the coding sessions for the free-response questions?

Number of NRCs		
Yes	No	No Response
32	10	1

F.6. By what date do you expect to have completed the coding?

	Number of NRCs
No date specified	1
April, 1995	2
May, 1995	5
June, 1995	12
July, 1995	11
August, 1995	5
September, 1995	1
October, 1995	2
November, 1995	2
December, 1995	2

The modal date was June, 1995.

F.7. Do you understand the procedure for coding the 10% reliability sample as explained in the *Guide to Checking, Coding, and Entering the TIMSS Data*?

Number of NRCs		
Yes	No	No Response
41	1	1

F.8. If no, please ask the NRC to explain.

The NRC who said "no" was yet to complete a study of the guide.

G. Data Entry and Transmittal

G.1. Do you plan to use your own staff or outside experts to enter the data from the achievement test booklets and questionnaires onto computer files?

		Number of NRCs
a.	own staff	13
b.	external data entry firm	9
c.	combination of the above	21
		n = 43

G.2. Have you selected the data entry staff?

NRCs		
Yes	No	No Response
25	17	1

G.3. If yes, have you conducted training sessions for the data entry staff?

Number of NRCs		
Yes	No	No Response
22	10	9

G.4. Do you plan to key enter a percentage of test booklets twice as a verification procedure?

Number of NRCs		
Yes	No	No Response
28	14	1

G.5. If yes, what percentage?

	Number of NRCs
No percentage specified	12
1 to 2 percent	2
3 to 5 percent	5
6 to 10 percent	12
11 to 15 percent	-
16 to 20 percent	-
21 to 25 percent	1
26 to 30 percent	-
31 percent or more	8

G.6. When do you plan to transmit the data to the IEA Data Processing Center in Hamburg, Germany?

	Number of NRCs
No date specified	2
May, 1995	3
June, 1995	3
July, 1995	7
August, 1995	13
September, 1995	8
October, 1995	1
November, 1995	1
December, 1995	3
February, 1996	1

The modal date was August, 1995

G.7. Have you established a secure storage area for the returned tests after coding and until the original documents can be discarded?

Number of NRCs		
Yes	No	Missing
39	4	-

H. Quality Assurance Sample

H.1. Have you selected your 10% quality assurance sample for your on-site classroom observations?

Number of NRCs		
Yes	No	No Response
20	23	-

H.2. If no, by what date do you plan to have this completed?

	Number of NRCs
No date specified	35
February, 1995	1
March, 1995	1
April, 1995	2
May, 1995	2
June, 1995	1
October, 1995	1

The modal date was May, 1995.

H.3. Who will do the Quality Assurance Classroom observations?

		Number of NRCs
a.	an external agency	4
b.	members of the NRC's staff	15
c.	a combination of the above	8
d.	other	5
e.	no response	11
		Total = 43

I. The NRC Report

I.1. Approximately by what date do you plan to send your NRC report to the TIMSS Study Center and the IEA Data Processing Center?

	Number of NRCs
No date specified	5
May, 1995	1
June, 1995	2
July, 1995	8
August, 1995	7
September, 1995	13
October, 1995	3
December, 1995	2
January, 1996	1
February, 1996	1

The modal date was September, 1995

I.2. Ask the NRC if he or she would like to comment on any aspect of the study, his or her role in it, problems that could have been avoided, etc.

Thirty four NRCs offered comments. These comments tended to focus upon difficulties experienced in the study.

		Number of Comments
a.	NRC expressed satisfaction with the study	4
b.	Project was too demanding	6
c.	There was a lack of resources (financial and staff)	6
d.	Lack of government support	3
e.	There was a need to reduce the demands on schools	5
f.	Translation and adaptation of TIMSS material required more time	4
g.	Manuals were too long, repetitive, and/or complex	4
h.	Questionnaires asked the wrong or were confusing	4
i.	Questionnaires were too long, in particular the teacher questionnaire	3
j.	Software problems were experienced	3

J. Collecting Materials From the NRC

J.1. As you discussed with the NRC in your pre-visit call, please request a copy of each of the following and indicate if it was obtained. Explain that you need the manuals and tracking forms (a-d) to assist in your school visits. Also, these documents as well as the test materials (e-i) and questionnaires (j-r) will be sent to the TIMSS Study Center to have on file as a possible reference source during the analysis process.

		Number of NRCs		
		Yes	No	No Response
a.	Test Administrator Manual	31	1	7
b.	School Coordinator Manual	26	1	9
c.	Student Tracking Forms for each class selected for observation	25	4	10
d.	Class Tracking Forms for each school selected for observation	22	4	14
e.	Population 1 Test booklets (8)	23	-	10
f.	Population 2 Test booklets (8)	34	-	9
g.	Population 3 Test booklets (up to 9)	18	3	9
h.	Population 1 Performance Assessment Tasks (12)	8	5	11
i.	Population 2 Performance Assessment Tasks (12)	17	5	11
j.	Population 1 School Questionnaire	23	-	11
k.	Population 2 School Questionnaire	35	-	8
l.	Population 3 School Questionnaire	18	3	9
m.	Population 1 Student Questionnaire	25	-	10
n.	Population 2 Student Questionnaire	35	-	8
o.	Population 3 Student Questionnaire	18	3	9
p.	Population 1 Teacher Questionnaire	23	-	10
q.	Population 2 Teacher Questionnaires (2)	34	-	10
r.	Translation Verification Report (obtain this from the NRC only if you did **not** receive a copy at the QA training session or from the Study Center)	24	4	12

APPENDIX F

Results of the Quality Assurance Monitors' Test Session Observations

A. Preliminary Activities of the Test Administrator

A total of 384 fully or partially completed Classroom Observation Records were received and processed. Two countries completed only section D, the interview with the School Coordinator, and so have no responses to sections A and C. In addition, one classroom observation did not have Section A completed. The number of classroom observation records included in this section is 373.

A.1. In your opinion, prior to the students' arrival, had the Test Administrator verified adequate supplies of the test books?

One classroom observation record did not record a response for this item.

		Pop 1 % of Responses	Pop 2 % of Responses	Pop 3 % of Responses	Combined % of Responses
a.	definitely yes	78	74	83	77
b.	probably yes	14	20	13	17
c.	probably not	1	2	1	2
d.	definitely not	6	4	3	4
e.	do not know	1	-	-	-
		Total = 81	Total = 222	Total = 69	Total = 372

A.2. Were all the seals INTACT on the test booklets prior to their distribution to the students?

Where seals were used, they were all intact. Two records were not completed, one each for population 1 and 3.

	% of Responses			
	Yes	No	Seals were not used	Total Responses
Pop 1	34	-	66	80
Pop 2	36	-	64	222
Pop 3	30	-	70	69
Combined	35	-	65	371

A.4. Does the student identification information on the test booklets and student questionnaires correspond with the *Student Tracking Form*?

Twelve classroom observation records did not have a response checked. Ten of these 12 came from a country which used a different Student Tracking Form (STF).

	% of Responses		
	Yes	No	Total
Pop 1	96	4	79
Pop 2	95	5	217
Pop 3	95	5	65
Combined	96	4	361

A.5. If no, please explain.

Seventeen classroom observation records had comments.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	The STF was not used	1	9	1	11
b.	The identification information on the test booklets did not correspond with the student tracking form	-	1	-	1
c.	Missing booklets replaced with spares	1	-	-	1
d.	No identification on the questionnaires	-	1	-	1
e.	Students were from different specialties	-	-	1	1
f.	An unexpected student showed up	-	1	-	1
g.	Two students were included in the group after the tracking form had been generated	-	-	1	1
		Total = 2	Total = 12	Total = 3	Total = 17

A.6. Did the Test Administrator have the correct version of the ADMINISTRATION SCRIPT for the assessment?

There was one Population 3 record sheet which was not checked for this item.

	% of Responses		
	Yes	No	Total
Pop 1	100	-	81
Pop 2	99	1	222
Pop 3	100	-	69
Combined	99	1	372

A.7. In your opinion, had the Test Administrator familiarized himself or herself with the SCRIPT prior to the testing?

There was one Population 3 record sheet which was not checked for this item.

		% of Responses			
		Pop 1	Pop 2	Pop 3	Combined
a.	definitely yes	70	73	80	74
b.	probably yes	23	18	13	19
c.	probably not	6	4	6	5
d.	definitely not	-	5	1	3
e.	do not know	-	-	-	-
		Total=81	Total=222	Total=69	Total=372

A.8. In your opinion, was there adequate seating space for the students to work without distractions?

Two classroom observation records, both from the same Population 3 country, did not have a check for a response category.

	% of Responses		
	Yes	No	Total
Pop 1	93	7	81
Pop 2	91	9	222
Pop 3	99	1	68
Combined	93	7	371

A.9. If no, please explain.

A total of 24 comments were recorded on the classroom observation records.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Students were closer than desired…	-	11	-	11
b.	Students were grouped………………	4	5	-	9
c.	The classrooms were small……….	2	2	-	4
		Total=6	Total=18	Total=0	Total=24

A.10. In your opinion, was there adequate room for the Test Administrator to move about during the testing to ensure that students were following directions correctly?

	% of Responses		
	Yes	No	Total
Pop 1	98	2	81
Pop 2	96	4	222
Pop 3	100	-	69
Combined	97	3	372

A.11. If no, please explain.

A total of 10 comments were recorded on the classroom observation forms.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	There was not enough space to walk around…………………………	1	7	-	8
b.	Too many students……………………	-	1	-	1
c.	Yes, between files of two students…………………………………	-	1	-	1
		Total = 1	Total = 9	Total = 0	Total = 10

A.12. Did the Test Administrator have a stop watch or timer for accurately timing the testing session(s)?

Four classroom observation records did not have a category checked. In some reports, the classroom observer indicated that an alternative timing device, for example a watch, was used by the test administrator. The classroom observer checked the "no" response category.

	% of Responses		
	Yes	No	Total
Pop 1	88	12	81
Pop 2	91	9	221
Pop 3.	90	10	67
Combined	90	10	369

A.13. Did the Test Administrator have an adequate supply of pencils and other necessary materials ready for the students?

Three classroom observation records did not have a response category checked.

	% of Responses		
	Yes	No	Total
Pop 1	86	14	81
Pop 2	73	27	222
Pop 3	75	25	67
Combined	77	23	370

A.14. Was there a wall clock visible for the students to check their timing during the testing?

Three classroom observation records did not have a response category checked.

	% of Responses		
	Yes	No	Total
Pop 1	36	64	81
Pop 2	20	80	220
Pop 3	36	64	69
Combined	26	74	370

B. Test Session Activities

There were 373 Classroom Observation Records with at least some responses for section B.

B.1. The Test Administrator should begin the session by reading aloud the sections of the ADMINISTRATOR'S SCRIPT entitled "Prepare the Students for the Test, Distribute Materials, and Begin Testing" through the instruction to "Start Working". Record the time elapsed for the Test Administrator to read these sections.

Nine classroom observation records did not have a response category checked.

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
Fewer than 5 minutes	3	4	14	6
5 to 9 minutes	19	33	45	32
10 to 14 minutes	51	39	26	39
15 to 19 minutes	23	16	9	16
20 to 24 minutes	4	5	5	5
25 or more minutes	-	3	1	2
Range	1 to 20	2 to 32	3 to 90	1 to 90
	Total = 79	Total = 218	Total = 67	Total =364

B.2. Did the Test Administrator follow the ADMINISTRATOR'S SCRIPT exactly in each of the following sections?

Four classroom observation records did not check a response.

a. Prepare the Students	% of Responses			
	Yes, No Changes	No, Minor Changes	No, Major Changes	Total
Pop 1	73	20	8	80
Pop 2	61	34	5	220
Pop 3	65	33	1	69
Combined	64	31	5	369

Nine classroom observation records did not record a response.

b. Distribute the Materials	% of Responses			
	Yes, No Changes	No, Minor Changes	No, Major Changes	Total
Pop 1	87	10	3	77
Pop 2	83	14	3	219
Pop 3	79	19	1	67
Combined	83	14	3	363

Seven classroom observation records did not record a response.

c. Begin Testing	% of Responses			
	Yes, No Changes	No, Minor Changes	No, Major Changes	Total
Pop 1	85	11	4	79
Pop 2.	81	15	4	219
Pop 3	79	19	1	68
Combined	82	15	3	366

B.3. If the Test Administrator made changes to the SCRIPT, how would you describe them?

In many cases, the Test Administrator adapted the narrative to suit the students. In all cases, the quality control monitors did not record changes which would have a detrimental effect on the test results.

a. Additions

	% of Responses		
	Yes	No	Total
Pop 1	68	32	22
Pop 2	73	28	80
Pop 3	59	41	27
Combined	69	31	129

b. Revisions

	% of Responses		
	Yes	No	Total
Pop 1	55	45	22
Pop 2	52	48	79
Pop 3	52	48	27
Combined	52	48	128

c. Deletions

	% of Responses		
	Yes	No	Total
Pop 1	45	55	20
Pop 2	52	48	67
Pop 3	57	43	28
Combined	52	48	115

B.4. Did the Test Administrator distribute test books one-at-a-time to each student?

There were four classroom observation records which did not record a response to this category. A total of 12 records indicated that the test books were not distributed one at a time.

	% of Responses		
	Yes	No	Total
Pop 1	100	-	80
Pop 2	99	1	220
Pop 3	86	14	69
Combined	97	3	369

B.5. If no, please explain.

Only 10 of the expected 12 classroom observation records gave a reason for the deviation from the prescribed distribution practice.

		# of Comments Pop 1	# of Comments Pop 3	# of Comments Pop 3	# of Comments Combined
a.	Booklets were placed on desks prior to student arrival	-	1	7	8
b.	Student picked up their test booklets from the TA's desk	-	1	1	2
		Total = 0	Total = 2	Total = 8	Total = 10

B.6. Did the Test Administrator distribute the test books according to the booklet assignments on the *Student Tracking Form*?

There were three classroom observation records which did not record a response. A total of 27 records showed a deviation from the prescribed procedure.

	% of Responses		
	Yes	No	Total
Pop 1	89	11	81
Pop 2	93	7	222
Pop 3	97	3	67
Combined	93	7	370

B.7. If no, please explain.

Only 21 of the 27 classroom observation records indicated a reason for the deviation from the prescribed procedure.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	Tracking form was not available or completed after the booklets were passed out	6	12	18
b.	Unexpected students showed up	1	-	1
c.	Absent student's booklet used for unexpected student	-	1	1
d.	One student received a booklet with a different number from the student tracking form	-	1	1
		Total=7	Total=14	Total=21

B.8. Did the Test Administrator record attendance correctly on the *Student Tracking Form*?

Seven classroom observation records did not record a response for this question. A total of 9 records indicated that attendance was recorded incorrectly.

	% of Responses		
	Yes	No	Total
Pop 1	98	2	81
Pop 2	98	2	217
Pop 3	97	3	68
Combined	98	2	366

B.9. Length of SESSION 1.

Nine observation records did not indicate the length of session 1.

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3
Fewer than 30 minutes	1	1	2
30 to 34 minutes	3	1	6
35 to 39 minutes	86	3	-
40 to 44 minutes	5	8	-
45 to 49 minutes	4	85	-
50 to 79 minutes	1	2	-
80 to 84 minutes	-	-	6
85 to 89 minutes	-	-	2
90 to 95 minutes	-	-	84
100 minutes or more	-	-	3
Range	29 to 60	6 to 74	24 to 136
	Total = 80	Total = 217	Total = 67

B.10. For the population being tested, did the total testing time for SESSION 1 equal the time allowed?

Nine classroom observation records did not record a response for this category. Four of these nine were from one country.

	% of Responses		
	Yes	No	Total
Pop 1	81	19	78
Pop 2	86	14	119
Pop 3	93	7	67
Combined	86	14	264

B.11. If no, please explain.

Of the 51 classroom observation records which showed a deviation from the time prescribed, 47 gave an explanation.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Most students finished early	2	11	3	15
b.	Less time given for unspecified reason	3	6	1	10
c.	1-5 minutes longer	5	3	-	9
d.	Time kept to 45 min. class period	-	7	-	7
e.	Time allowed printed on test was 46 min.	3	-	-	3
f..	Mis-timed by TA (29 instead of 37 minutes)	1	-	-	1
g.	Students allowed to proceed with part 2 when complete with part 1	-	1	-	1
h.	Students were given 23 minutes more than the time that should have been allotted	1	-	-	1
		Total=15	Total=28	Total=4	Total=47

B.12. Did the Test Administrator announce "you have 10 minutes left" prior to the end of SESSION 1?

Three classroom observation records did not indicate a response for this question.

	% of Responses		
	Yes	No	Total
Pop 1	89	11	80
Pop 2	86	14	221
Pop 3	84	16	69
Combined	86	14	370

B.13. Were there any other "time remaining" announcements made during SESSION 1?

Eight classroom observation records did not show a response to this item.

	% of Responses		
	Yes	No	Total
Pop 1	20	80	80
Pop 2	24	76	116
Pop 3	28	72	69
Combined	24	76	365

B.14. If yes, please explain.

Of the 88 classroom observation records which indicated an additional announcement, 79 offered a description of the announcement.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	An additional announcement was made	11	33	8	52
b.	2-3 additional announcements were made	2	12	6	20
c.	Time given in response to student questions	1	3	2	6
d.	Time was marked on a blackboard	-	-	1	1
		Total=14	Total=48	Total=17	Total=79

B.15. When the Test Administrator ended SESSION 1, how well did the students comply with the instruction to "Stop work"?

Nine classroom observation records did not show a response to this item. Of these 9, 4 were from one country.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3
a.	very well, all students stopped work	80	74	77
b.	well, almost all students stopped work	17	22	22
c.	fairly well, some students did not stop	3	2	-
d.	not well at all, many students did not stop	-	2	1
		Total = 81	Total = 215	Total = 68

B.16. At the end of SESSION 1, did the Test Administrator collect the test books one-at-a-time from each student?

There were a total of 6 classroom observation records which did not indicate a response.

	% of Responses		
	Yes	No	Total
Pop 1	51	49	80
Pop 2	41	59	221
Pop 3	82	18	76
Combined	51	49	367

B.17. If no, please explain.

While 180 classroom observation records indicated that test booklets were not collected in the prescribed manner, only 145 explanations were given. In some reports, mention was made of the use of envelopes. In these cases, students placed the test booklets back into the envelopes before the break.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a .	Test books were collected as students finished	-	1	-	1
b.	Booklets were not collected because of shortened or eliminated break	5	24	-	29
c.	Test books remained on students desks for duration of the break	29	77	6	112
d.	Booklets were passed forward	-	1	-	1
e.	Miscellaneous	-	-	2	2
		Total=33	Total=98	Total=8	Total=145

B.18. Length of the BREAK.

There were 303 classroom observation records for populations 1 and 2. A total of 35 forms did not record a response to this item.

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Combined
No time specified	5	9	8
9 or fewer minutes	18	17	17
10 to 14 minutes	23	29	27
15 to 19 minutes	8	18	15
20 to 25 minutes	29	18	21
25 to 29 minutes	7	3	5
30 or more minutes	10	5	7
Range	0 to 43	0 to 39	0 to 43
	Total = 77	Total = 191	Total = 268

B.19. Was the total time for the BREAK between SESSION 1 and SESSION 2 equal to 20 minutes?

	% of Responses		
	Yes	No	Total
Pop 1	28	72	79
Pop 2	16	84	203
Combined	19	81	282

B.20. Was the BREAK conducted as directed in the SCRIPT?

A little over half of the breaks were conducted according to the script. Twenty one classroom observation records did not indicate a response to this item.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Combined
a.	exactly as directed	63	53	56
b.	nearly the same as directed	27	28	28
c.	somewhat differently	9	14	12
d.	not well at all, many students did not stop	1	4	4
		Total = 81	Total = 201	Total = 282

B.21. If not "exactly as directed," please explain.

Where the break was not "exactly as directed", some test administrators explained the disparity as following the school routine or by pointing out that the directions could reasonably be interpreted as guidelines. Many classroom observers simply noted whether or nor the break was according to the script or did not offer an explanation. Of the 124 classrooms which did not follow the script exactly, there were 103 comments offered.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	The script was not consulted when the break was conducted	-	2	2
b.	The break time was longer than specified in the script	6	15	21
c.	The break time was shorter than the script specified	9	26	34
d.	No break was conducted	7	31	38
e.	Potential for security problems during the break	1	-	1
f.	Unspecified time difference	2	4	6
		Total=25	Total=78	Total=103

B.22. Record the restart time for the Test Administrator to begin SESSION 2 once all the students have returned from the BREAK. This time period begins when the Test Administrator is ready to distribute the test books for SESSION 2 and reads the instructions through to the instruction to "Start Working".

Restart time:	% of Responses Pop 1	% of Responses Pop 2	% of Responses Combined
No time specified	5	11	9
4 or fewer minutes	45	57	54
5 to 9 minutes	43	28	32
10 to 14 minutes	3	2	2
15 to 19 minutes	1	-	1
20 to 24 minutes	-	-	-
25 or more minutes	3	2	2
Range	1 to 32	1 to 60	1 to 60
	Total = 77	Total = 203	Total = 280

B.23. Was the time spent to restart the testing with SESSION 2, 5 minutes?

	% of Responses		
	Yes	No	Total
Pop 1	44	56	80
Pop 2	34	66	211
Combined	37	63	291

B.24. If no, please explain.

Most of the classroom observation records comment on the time span, rather than offering an explanation. In cases where an explanation is give, the comments include "children had no questions", " students were told to start work again" and "students embarked immediately on the second part of the test".

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	Less time was required to restart testing than specified in the script..	10	16	26
b.	More time was required to restart testing than specified in the script	6	10	16
c.	Restart time was not needed	5	30	35
d.	Some clarification for a late student took longer made the restart take longer than the specified time	1	-	1
e.	Comment suggests that restart time fell in specified range	9	47	56
		Total=31	Total=103	Total=134

B.25. Length of SESSION 2.

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Combined
Fewer than 20 minutes	1	-	-
20 to 29 minutes	85	1	24
30 to 39 minutes	6	8	8
40 to 49 minutes	4	89	66
50 to 59 minutes	1	1	1
60 or more minutes	3	1	1
Range	19 to 80	29 to 106	19 to 106
	Total = 80	Total = 214	Total = 294

B.26. Was the total time for testing in SESSION 2 correct as indicated in the ADMINISTRATORS' SCRIPT?

	% of Responses		
	Yes	No	Total
Pop 1	75	25	80
Pop 2	81	19	217
Combined	79	21	297

B.27. If no, please explain.

Classroom observers tended to focus upon the time span rather than recording explanations for the disparity.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	Less time was required for testing in session 2 than given	10	22	21
b.	The test was mistimed by the TA	1	-	1
c.	1-5 minutes longer was given than allowed for	5	11	16
d.	Testing time was shorter than specified time for an unspecified reason	1	-	1
e.	A slow student was given additional time	2	-	2
f.	Unspecified additional time	1	-	1
		Total=20	Total=32	Total=52

B.28. Did the Test Administrator announce "you have 10 minutes left" prior to the end of SESSION 2?

	% of Responses		
	Yes	No	Total
Pop 1	80	20	81
Pop 2	81	19	218
Combined	81	19	299

B.29. Were there any other "time remaining" announcements made during SESSION 2?

	% of Responses		
	Yes	No	Total
Pop 1	16	84	80
Pop 2	21	79	215
Combined	20	80	295

B.30. If yes, please explain.

Frequently the test administrator made several announcements. In some cases the announcement was seen as redundant since most students had already finished.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	An additional announcement was made	7	31	38
b.	More than one additional announcements were made	3	8	11
e.	Announcement made about break and SQ	-	3	3
d.	TA responded to students questions	-	2	2
c.	Time was written on the black board	-	1	1
		Total=10	Total=45	Total=55

B.31. When the Test Administrator ended SESSION 2, how well did the students comply with the instruction to "Stop Work"?

		% of Responses Pop 1	% of Responses Pop 2
a.	very well, all students stopped work	84	80
b.	well, almost all students stopped work	16	16
c.	fairly well, some students did not stop	-	3
d.	not well at all, many students did not stop	-	-
		Total = 79	Total = 213

B.32. At the end of SESSION 2, did the Test Administrator collect the test books one-at-a-time from each student?

	% of Responses		
	Yes	No	Total
Pop 1	89	11	80
Pop 2	78	22	218
Combined	81	19	298

B.33. If no, please explain.

Classroom observers frequently reported that the booklets were left on the desk to be collected after the students had left the room. Explanations included " the administrator stayed in the room after students had left", " the administrator locked the room" or" there was no break".

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	Test Booklets were left on students desks	7	25	32
b.	Booklets were not collected. TA went directly on to the student questionnaire	1	2	3
c.	Test Booklets were passed to the front of the room	-	3	3
		Total=8	Total=30	Total=38

B.34. When the Test Administrator read the SCRIPT for the end of the testing SESSION 2, did the Test Administrator announce a BREAK to be followed by the *Student Questionnaire*?

	% of Responses		
	Yes	No	Total
Pop 1	71	29	79
Pop 2	65	35	188
Combined	67	33	267

B.35. If no, please explain.

Classroom observers recorded a range of explanations for deviations from the script. These included that the manual did not recommend a break, there was no script in the version being used, or the test administrator offered a more concise interpretation of the script.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	No break was provided	6	27	33
b.	The student questionnaire was administered at a different time	5	22	27
c.	Manual not used	-	7	7
d.	Questionnaires distributed before the test began	4	-	4
e.	Questionnaire was not announced	-	3	3
f.	Questionnaire not administered	1	-	1
		Total=16	Total=59	Total=75

B.36. How accurately did the Test Administrator read the SCRIPT to end the testing and signal a BREAK?

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Combined
a.	verbatim: no changes	64	63	63
b.	some minor changes	25	27	26
c.	major changes	12	11	11
		Total = 69	Total = 160	Total = 229

B.37. If the Test Administrator made changes to the SCRIPT, how would you describe them?

a. additions

	% of Responses Yes	% of Responses No	Total
Pop 1.	30	70	30
Pop 2	31	69	47
Combined	30	70	77

b. some minor changes

	% of Responses Yes	% of Responses No	Total
Pop 1.	41	59	32
Pop 2	56	44	55
Combined	51	49	87

c. omissions

	% of Responses Yes	% of Responses No	Total
Pop 1.	35	65	34
Pop 2	46	54	57
Combined	42	58	91

B.38. Please record how long the BREAK was between the end of the testing sessions and the distribution of the *Student Questionnaire.*

	% of Responses Pop 1	% of Responses Pop 2
No time specified	3	5
Less than 10 minutes	26	33
10 to 19 minutes	32	42
20 to 29 minutes	31	14
30 to 39 minutes	3	3
40 or more minutes	5	2
Range	1 to 73	1 to 155
	Total = 62	Total = 154

B.39. How did the actual BREAK time compare to the recommended time in the SCRIPT?

		% of Responses Pop 1	% of Responses Pop 2
a.	exactly the same	48	51
b.	it was longer	30	22
c.	it was shorter	21	27
		Total = 56	Total = 141

B.40. If not "exactly the same," how much longer or shorter?

	% of Responses Pop 1	% of Responses Pop 2
No time specified	4	-
1 to 9 minutes shorter	4	5
Fewer than 5 minutes longer	42	39
5 to 9 minutes longer	33	30
10 to 19 minutes longer	8	23
20 or more minutes longer	9	3
Range	-9 to 30	-2 to 100
	Total = 24	Total = 61

B.41. Was the BREAK conducted as directed in the SCRIPT?

		% of Responses Pop 1	% of Responses Pop 2
a.	exactly as directed	78	65
b.	nearly the same as directed	12	15
c.	somewhat differently	5	9
d.	very differently	5	11
		Total = 58	Total = 149

B.42. If not "exactly as directed," please explain.

In several classroom observation records it was mentioned that there was no directions in the script or that the test administrators adapted the script according to directions given by the NRC or external agency, or students did not leave the classroom.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	No break was conducted	2	4	6
b.	The student questionnaire was administered in a different classroom	-	1	1
c.	The length of the break was shortened	-	12	12
d.	The length of the break was increased	5	5	10
e.	Students remained at their seat for the duration of the break	-	7	7
f.	The student questionnaire was completed on a different day	1	9	10
g	No directions about the break were specified in the TA manual	1	4	5
h.	Unspecified difference	-	2	2
		Total=9	Total=44	Total=53

B.43. At the end of the BREAK, did the Test Administrator distribute the *Student Questionnaires* and give directions as specified in the SCRIPT?

	% of Responses		
	Yes	No	Total
Pop 1	73	27	60
Pop 2	65	35	168
Combined	67	33	228

B.44. If no, please explain.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	Student questionnaires were included within the test booklet	5	7	12
b.	Student questionnaires were scheduled to be completed at another time	5	9	14
c.	Questionnaires were distributed prior to student's arrival	3	13	16
d.	TA distributed questionnaires and instructions to each student individually	1	1	2
e.	Some omissions were made when the instructions were read	-	1	1
f.	"Names were indicated on the booklet	-	1	1
g.	No break	-	2	2
h.	No directions were addressed to the entire group	-	1	1
i.	TA went very quickly through the instructions	-	1	1
j.	Almost exactly	2	3	5
		Total=16	Total=39	Total=55

B.45. Length of time for administration of the *Student Questionnaire*.

	% of Responses Pop 1	% of Responses Pop 2
19 or fewer minutes	3	8
20 to 29 minutes	39	30
30 to 39 minutes	32	38
40 to 49 minutes	16	18
50 to 59 minutes	7	3
60 or more minutes	3	3
Range	2 to 62	2 to 95
	Total = 57	Total = 154

B.46. How does the total time allocated for the administration of the *Student Questionnaire* compare to the time specified in the SCRIPT?

		% of Responses Pop 1	% of Responses Pop 2
a.	exactly the same	35	39
b.	it was longer	63	59
c.	it was shorter	2	3
		Total = 57	Total = 157

B.47. If not "exactly the same," how much longer or shorter?

	% of Responses Pop 1	% of Responses Pop 2
10 or more minutes shorter	-	1
9 or fewer minutes shorter	-	2
9 or fewer minutes longer	31	33
10 to 19 minutes longer	36	39
20 to 29 minutes longer	30	19
30 or more minutes longer	3	6
Range	1 to 30	-10 to 45
	Total = 178	Total = 93

B.48. Did the students ask for additional time to complete the questionnaire?

	% of Responses		
	Yes	No	Total
Pop 1	61	39	59
Pop 2	57	43	160
Combined	58	42	219

B.49. How much additional time was given to complete the *Student Questionnaire*?

	% of Responses Pop 1	% of Responses Pop 2
Fewer than 5 minutes	6	7
5 to 9 minutes	19	28
10 to 14 minutes	19	33
15 to 19 minutes	17	11
20 to 24 minutes	28	17
25 or more minutes	11	4
Range	2 to 30	2 to 45
	Total = 36	Total = 90

B.50. At the end of the session, prior to dismissing the students, did the Test Administrator thank the students for participating in the study?

	% of Responses		
	Yes	No	Total
Pop 1	76	24	63
Pop 2	81	19	178
Combined	80	20	241

B.51. In your opinion, how orderly was the dismissal of the students?

Twenty classroom observation records for population 1 and a total of 52 for population 2 did not record a response for this item..

		% of Responses Pop 1	% of Responses Pop 2
a.	very orderly	74	58
b.	somewhat orderly	20	36
c.	not orderly at all	6	6
		Total = 65	Total = 176

B.69. When the Test Administrator read the SCRIPT to end the 90 minute testing session, did the Test Administrator announce a BREAK?

There were 9 classroom observation records which did not indicate a response to this item.

% of Responses		
Yes	No	Total
64	36	56

B.70. If no, please explain.

Of the 22 classroom observation records which indicated that the break was not announced, 12 provided an explanation.

		# of Comments Pop 3
a.	No break was given	9
b.	Questionnaire had been scheduled for another time; therefore no break was necessary	2
c.	Most students had already taken a break as they finished the test early	1
		Total=12

B.71. How accurately did the Test Administrator read the SCRIPT to end the testing and signal a BREAK?

		% of Responses
a.	verbatim: no changes	52
b.	some minor changes	39
c.	major changes	9
		Total = 56

B.72. If there were changes, how would you describe them?

Several classroom observation records indicated that test administrators made multiple changes.

		% of Responses		
		Yes	No	Total
a.	additions	14	86	14
b.	some minor changes	63	37	19
c.	deletions	50	50	18

B.73. Please record how long the BREAK was between the end of the testing sessions and the distribution of the *Student Questionnaire.*

	% of Responses Pop 3
Fewer than 10 minutes	51
10 to 14 minutes	13
15 to 19 minutes	16
20 to 24 minutes	10
25 or more minutes	10
Range	1 to 30
	Total = 31

B.74. How did the actual BREAK time compare to the recommended time in the SCRIPT?

		% of Responses
a.	exactly the same	58
b.	it was longer	9
c.	it was shorter	33
		Total = 41

B.75 If not "exactly the same," how much longer or shorter?

	% of Responses
No time specified	8
21 to 30 minutes shorter	8
11 to 20 minutes shorter	15
1 to 10 minutes shorter	39
1 to 10 minutes longer	23
50 minutes longer	8
Range	-30 to 50
	Total = 33

B.76. Was the BREAK conducted as directed in the SCRIPT?

		% of Responses
a.	exactly as directed	66
b.	nearly the same as directed	11
c.	somewhat differently	2
d.	very differently	20
		Total = 44

B.77. If not "exactly as directed," please explain.

		# of Comments Pop 3
a.	No break was given	9
b.	No directions in the script concerning break time -	1
		Total=10

B.78. At the end of the BREAK, did the Test Administrator distribute the *Student Questionnaires* and give directions as specified in the SCRIPT?

Fifteen classroom observation reports did not record a response for this item.

% of Responses		
Yes	No	Total
71	29	55

B.79. If no, please explain.

There were 16 classroom observation records which indicated that there was a deviation from the script.

		Number of Comments
a.	Administration of student questionnaire was scheduled for another time	2
b.	Students had been given the questionnaire at the same time as the test	14
c.	"No break was given"	1
d.	The questionnaire was distributed but the directions specified in the script, were not given	1
e.	Affirmative explanation	1
f.	The teacher answered many questions	1
g.	The student questionnaire was sent home	1
		Total = 21

B.80. Length of time for administration of the *Student Questionnaire.*

	% of Responses
10 to 14 minutes	2
15 to 19 minutes	17
20 to 24 minutes	13
25 to 29 minutes	26
30 to 34 minutes	32
35 to 39 minutes	4
40 or more minutes	6
Range	12 to 65
	Total = 53

B.81. How does the total time allocated for the administration of the *Student Questionnaire* compare to the time specified in the SCRIPT?

		% of Responses
a.	exactly the same	34
b.	it was longer	26
c.	it was shorter	40
		Total = 50

B.82. If not "exactly the same," how much longer or shorter?

	% of Responses
11 to 15 minutes shorter	7
6 to 10 minutes shorter	14
1 to 5 minutes shorter	38
1 to 5 minutes longer	28
6 to 10 minutes longer	7
21 to 30 minutes longer	3
31 to 35 minute longer	7
Range	-14 to 35
	Total = 29

B.83. Did the students ask for additional time to complete the questionnaire?

% of Responses		
Yes	No	Total
23	75	66

B.84. How much additional time was given to complete the *Student Questionnaire*?

	% of Responses
No time specified	11
1 to 5 minutes	67
6 to 10 minutes	6
11 to 15 minutes	-
16 or more minutes	17
Range	1 to 35
	Total = 18

B.85. At the end of the session, prior to dismissing the students, did the Test Administrator thank the students for participating in the study?

Ten classroom observation records did not show a response.

% of Responses		
Yes	No	Total
82	18	60

B.86. In your opinion, how orderly was the dismissal of the students?

There were 9 classroom observation records which did not record a response. Three of these nine represent all the classroom observations from one country, while a further two came from another country.

		% of Responses
a.	very orderly	75
b.	somewhat orderly	20
c.	not orderly at all	5
		Total = 61

C. Summary Observations

In addition to the two countries did not complete section C of the Classroom Observation Record, three of the Classroom Observation Records that were submitted to the Study Center omitted the summary observations section. The total number of sessions for which summary observations are available is 371.

C.1. To what extent would you describe the students as orderly and cooperative?

Two of the sessions observed provided no response to this question.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	extremely orderly and cooperative	80	56	77	65
b.	moderately orderly and cooperative	19	36	20	29
c.	somewhat orderly and cooperative	1	7	1	5
d.	hardly cooperative at all	-	1	1	1
		Total = 80	Total = 220	Total = 69	Total = 369

C.1. Comments:

Only 50 sessions provided comments along with this item.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Positive comment	8	8	7	23
b.	Some behavior problems or minor disruptions were noted	5	9	5	19
c.	There was some confusion in the beginning with regard to seating and booklets	-	1	1	2
d.	Comments not directly related to the question	1	3	2	6
		Total=14	Total=21	Total=15	Total=50

C.2. If the students were not cooperative and orderly, did the Test Administrator make an effort to control the students and the situation?

Students in many sessions were described as orderly and cooperative. Only 110 observation records had responses that indicated that any students were uncooperative.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	definitely yes	80	70	64	71
b.	some effort was made	19	26	29	25
c.	hardly any effort was made	-	4	7	4
		Total = 16	Total = 80	Total = 14	Total =110

C.3. During the testing sessions did the Test Administrator walk around the room to be sure students were working on the correct section of the test and/or behaving properly?

	% of Responses		
	Yes	No	Total
Pop 1	94	6	80
Pop 2	94	6	222
Pop 3	96	4	69
Combined	94	6	371

C.4. If no, please explain.

Only 18 of the 21 session answering 'no' to question C3 provided comments.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	TA did not walk around	1	8	2	11
c.	Not enough space to walk around	1	2	-	3
d.	Not very frequently	2	-	-	2
e.	TA completed their questionnaire	-	2	-	2
		Total=4	Total=12	Total=2	Total=18

C.5. In your opinion, did the Test Administrator address students' questions appropriately?

In approximately 16 sessions, quality control monitors felt that they could not assess the TA's response to student questions because either no questions were asked or the QCM was not in a position to hear how the teacher responded to individual questions. Two observations left the item blank with no further explanation.

	% of Responses		
	Yes	No	Total
Pop 1	96	4	78
Pop 2	96	4	210
Pop 3	100	-	65
Combined	97	3	353

C.6. If no, please explain.

Ten of the 12 observation records responding 'no' provided comments.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Some questions were answered	2	3	-	5
c.	The QCM felt that the TA was too reluctant to answer student questions	-	3	-	3
d.	The TA was particularly nervous	-	1	-	1
b.	No questions were answered	-	1	-	1
		Total=2	Total=8	Total=0	Total=10

C.7. Did you see any evidence of students attempting to cheat on the tests (e.g., by copying from a neighbor)?

Two observation records had no response for this item. The QCM from one of these indicated that he saw the potential for cheating but did not feel that he actually observed students attempting to cheat.

	% of Responses		
	Yes	No	Total
Pop 1	14	86	79
Pop 2	14	86	221
Pop 3	9	91	69
Combined	13	87	369

C.8. If yes, please explain.

Forty-three of the 48 sessions in which QCMs saw some evidence of cheating provided comments.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Talking and copying was noted	1	15	4	20
b.	Some students tried to talk with their neighbors	3	8	-	11
c.	Attempts were stopped	4	2	3	9
d.	Some students were noticed "glancing" at others papers	1	1	-	2
e.	QCM acknowledged that attempts took place but did not specify outcome	-	1	-	1
		Total=9	Total=27	Total=7	Total=43

C.9. Were any defective test books detected and replaced before the testing began?

	% of Responses		
	Yes	No	Total
Pop 1	4	96	80
Pop 2	8	92	222
Pop 3	1	99	69
Combined	6	94	371

C.10. Were any defective test books detected and replaced after the testing began?

Three observation records provided no response to this question.

	% of Responses		
	Yes	No	Total
Pop 1	8	92	79
Pop 2	6	94	221
Pop 3	3	97	68
Combined	6	94	368

C.11. If any defective test books were replaced, did the Test Administrator replace them appropriately?

	% of Responses		
	Yes	No	Total
Pop 1	100	-	6
Pop 2	74	26	23
Pop 3	100	-	3
Combined	81	19	32

C.12. If no, please explain.

Only six of the 32 classroom observation records in which defective booklets were identified indicated that established procedures were not followed. These 6 provided the following comments.

		# of Comments Pop 2
a.	The problem did not necessitate replacing the book	3
b.	No spares to rely on	1
c.	Problem discovered at end of session	1
d.	Given the booklet of an absent student	1
		Total=6

C.13. Were any late students admitted to the testing room?

Five of the observation records had no response for this question.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	no, there were no late students	93	90	74	88
b.	no, they were not admitted	-	2	4	2
c.	yes, but before the testing began	5	4	12	5
d.	yes, after the testing began	3	4	10	5
		Total = 80	Total = 218	Total = 68	Total = 366

C.14. Did any students refuse to take the test either prior to the testing or during the testing?

Six observation records had no response checked for this item. Of these, 2 had comments indicating that a student was reluctant to work but was persuaded to try. Two had comments indicating that a number of students did not attend, and 2 were simply left blank.

	% of Responses		
	Yes	No	Total
Pop 1	1	99	78
Pop 2	2	98	220
Pop 3	6	94	67
Combined	3	97	365

C.15. If yes, how many students refused to complete the testing?

Seven of the 10 observation records indicating that students refused to take the test responded to item C15.

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
1 to 5 students	100	75	100	86
6 to 10 students	-	-	-	-
11 to 15 students	-	-	-	-
16 to 20 students	-	-	-	-
21 to 25 students	-	25	-	14
Range	1 to 1	1 to 23	1 to 2	1 to 23
	Total = 1	Total = 4	Total = 2	Total = 7

C.16. If a student refused, did the Test Administrator accurately follow the instructions for excusing the student (collect the test book and record the incident on the *Student Tracking Form*)?

One of the 10 observations in which at least one student refused to take the test left this item blank on the observation record.

	% of Responses		
	Yes	No	Total
Pop 1	100	-	1
Pop 2	40	60	5
Pop 3	67	33	3
Combined	56	44	9

C.16. Comments:

Six of the 10 observations that checked 'yes' to item 14 provided comments.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Reluctant students were given a motivational talk	-	2	-	2
b.	Student excluded because she was studying courses for the mentally handicapped	1	-	-	1
c.	While the students were somewhat fearful, they worked hard	-	1	-	1
d.	Two students stopped working without informing the TA	-	1	-	1
e.	Student left the administration early	-	-	1	1
		Total=1	Total=4	Total=1	Total=6

C.17. Did any students leave the room for an "emergency" during the testing?

Seven of the 371 classroom observation records for which summary observations were obtained did not respond to this question.

	% of Responses		
	Yes	No	Total
Pop 1	11	89	79
Pop 2	11	89	218
Pop 3	33	67	67
Combined	15	85	364

C.18. If yes, did the Test Administrator address the situation appropriately (collect the book, and if re-admitted, return the test book and record time out of the room on the test book)?

Six of those responding 'yes' to question C17 did not provide a response.

	% of Responses		
	Yes	No	Total
Pop 1	75	25	8
Pop 2	62	38	21
Pop 3	45	55	19
Combined	57	43	49

C.18. Comments:

Thirty-five of the classroom observation records provided comments with C18. Several of the comments fell into more than one category.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Student left the room for a short period of time	4	5	5	14
b.	Booklets left on students desks	4	4	5	13
c.	Students left after completing the test	-	3	4	7
d.	Time out was not recorded	-	3	2	5
e.	Minor emergencies occurred (Nose Bleed, Stomach ache, dental operation)	-	3	-	3
f.	Notations were made on the books	-	-	1	1
g.	QCM noted that procedures were not followed but did not feel that this caused a problem	1	-	-	1
		Total=9	Total=18	Total=17	Total=44

C.19. In general, how would you describe the overall quality of the testing session?

Four classroom observations did not have responses for this question.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	excellent	58	37	68	47
b.	very good	28	41	19	34
c.	good	14	14	9	13
d.	fair	1	5	4	4
e.	poor	-	4	-	2
		Total = 80	Total = 219	Total = 68	Total = 367

C.20. Please feel free to comment on any aspect of the assessment you wish to note.

164 of the test administrators provided comments. These are summarized below:

Subject of Comment	Nature of Comment	Total
Test Administrator	Well organized, prepared, conducted the test according to guidelines, excellent supervision of students	38
	Poorly organized, unmotivated and uncooperative, inadequate supervision of students	16
Students	Well disciplined, motivated, challenged by the test	25
	Anxious, found the test level too high/too difficult	2
Test	Excellent, well presented and conducted	2
	Time allocated to test inappropriate	26
	Students tired at end of sessions	6
	Language inappropriate, content mismatch with curriculum	17
Test Conditions	Poor	4

D. Interview With the School Coordinator

All of the 384 Classroom Observation Records returned to the Study Center have collected at least some information for section D.

D.1. Before I leave, I would just like to ask you a few questions about your experiences with the TIMSS testing. Overall, how would you say the session went? Would you say it went very well, satisfactorily, or unsatisfactorily?

Nine Classroom Observation Records have no response for this question.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	Very well, no problems	78	64	78	70
b.	Satisfactorily, few problems	21	34	19	29
c.	Unsatisfactorily, many problems	1	1	3	2
		Total = 85	Total = 221	Total = 69	Total = 375

D.1. Comments

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Difficult time of year	4	8	4	16
b.	Insufficient administration time	1	9	3	13
c.	Problems with booklets, questionnaires	3	5	-	8
e.	Minor staff problems\concerns	1	5	1	7
d.	Generally positive comment	-	6	-	6
f.	Student\students did not attend	-	-	4	4
g.	Time too generous	1	1	1	3
h.	Behavior problems or minor disruptions	1	2	-	3
i.	Test was found to be difficult	-	1	1	2
j.	Not enough chairs, spare rooms available	-	2	-	2
k.	School would like to see results	-	1	-	1
l.	Performance assessment difficult to arrange	-	1	-	1
m.	Some students could not stay because of other commitments	-	-	1	1
		Total=11	Total=42	Total=14	Total=67

211

D.2. Overall, how would you rate the attitude of the other school staff members towards the TIMSS testing? Would you say that it was positive, neutral, or negative?

Nine observation records did not provide a response to this item.

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	Positive	67	73	69	71
b.	Neutral	25	24	29	25
c.	Negative	8	3	1	4
		Total = 85	Total = 222	Total = 68	Total = 375

D.2. Comments

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Generally positive comment	5	24	12	41
b.	Difficult time of year	2	8	5	15
c.	Negative attitude displayed by some staff	3	5	5	13
d.	Time consuming\lots of work	2	7	-	9
e.	Most staff had little or no information about the testing	-	1	2	6
f.	Some neutral	1	3	-	4
g.	Concerned about test results	2	1	1	4
h.	Test questions were difficult	2	1	-	3
i.	Want feedback	-	3	-	3
j.	Problems with questionnaires	1	1	-	3
k	Miscellaneous	1	5	1	7
		Total=19	Total=60	Total=29	Total=108

D.3. Prior to the test day did you have time to check your shipment of materials from your TIMSS National Coordinator?

Fourteen of the classroom observation records did not provide a response for this item.

	% of Responses		
	Yes	No	Total
Pop 1	88	12	84
Pop 2	86	14	219
Pop 3	91	9	67
Combined	87	13	370

D.4. If yes, how long before the test date?

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
No time specified	7	15	20	14
1 day	12	25	7	19
2 to 5 days	18	22	15	20
6 to 10 days	31	15	31	22
11 to 15 days	22	14	15	17
16 to 20 days	-	1	5	1
21 or more days	11	7	8	8
Range	1 to 30	1 to 75	1 to 21	1 to 75
	Total = 74	Total = 188	Total = 61	Total = 323

D.5. Did you receive the correct shipment of the following items?

A large number of QCMs did not check one or more of the provided responses for this item.

		% of Responses		
		Yes	No	Total
a.	Test booklets	99	1	368
b.	Test Administrator Manual	100	-	359
c.	School Coordinator Manual	98	2	324
d.	Student Tracking Forms	93	7	365
e.	Student Questionnaires	98	2	368
f.	Teacher Questionnaires	91	9	335
g.	School Questionnaire	99	1	368
h.	Test Administration Form	92	8	350
i.	Teacher Tracking Form	70	30	310
j.	Student-Teacher Linkage Form (if applicable)	26	74	197
k.	Envelopes or boxes addressed to the national center for the purpose of returning the materials after the assessment	71	29	340

D.6. Was the National Coordinator responsive to your questions or concerns?

This item was left blank on 53 of the observation records. Several of the QCMs who left this item blank indicated that the School Coordinator had had not required contact with the National Coordinator.

	% of Responses		
	Yes	No	Total
Pop 1	96	4	69
Pop 2	94	6	198
Pop 3	89	11	64
Combined	93	7	331

D.6. Comments:

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Did not have any questions or concerns	11	21	6	38
b.	Generally positive comment	11	18	9	38
c.	Very little contact	2	6	-	8
d.	Training very short	1	2	-	3
e.	Contact was with regional coordinator	-	1	2	3
f.	The coordinator received previous training	-	3	-	3
g.	The National Coordinator delegated the responsibilities to other staff members	-	3	-	3
h.	Miscellaneous	1	1	2	4
		Total=26	Total=55	Total=19	Total=100

D.7. Were you able to collect completed *Teacher Questionnaires* prior to the test administration?

Sixty classroom observation records provided no response for this question.

	% of Responses		
	Yes	No	Total
Pop 1	42	58	78
Pop 2	42	58	218
Pop 3	14	86	29
Combined	40	60	325

D.7. Comments:

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Not complete/collected at time of interview	13	41	3	57
b.	Questionnaires have been collected	3	28	1	32
c.	Some collected at the time of the interview	4	11	-	15
d.	Teachers disliked doing or had problems with the questionnaire	3	10	-	13
e.	Not done in this school	-	1	5	6
f.	Given out after testing	-	2	-	2
		Total=23	Total=97	Total=9	Total=129

D.8. It was expected that the *Teacher Questionnaires* would require about 60 minutes to complete. In your opinion, was that estimate correct?

		% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
a.	yes	52	31	42	36
b.	no, it took longer	46	65	50	60
c.	no, it took less time	2	4	8	4
		Total = 56	Total = 185	Total = 12	Total = 253

D.9. How much longer or shorter?

	% of Responses Pop 1	% of Responses Pop 2	% of Responses Pop 3	% of Responses Combined
30 or more minutes shorter	-	2	14	2
15 to 29 minutes shorter	4	3	-	3
1 to 14 minutes shorter	-	-	-	-
1 to 14 minutes longer	16	9	-	10
15 to 29 minutes longer	16	11	-	11
30 to 44 minutes longer	20	33	57	32
45 to 59 minutes longer	-	3	14	3
60 or more minutes longer.	44	40	14	40
Range	-20 to 90	-50 to 120	-30 to 60	-50 to 120
	Total = 25	Total = 114	Total = 7	Total = 146

D.10. Where were the testing materials stored prior to the test administration?

		# of Responses
a.	Office of the principal or assistant principal	88
b.	An off site location	72
c.	Locked filing cabinet; locked cupboard or vault	40
d.	Administration office/storage room	36
e.	Teacher's office	27
f.	Office of the school coordinator\department head	21
g.	In the classroom	8
h.	Conference room	8
i.	"not applied"	9
		Total = 309

D.11. Were you satisfied with the accommodations (testing room) you were able to arrange for the testing?

Four of the classroom observation records provided no response to this item.

	% of Responses		
	Yes	No	Total
Pop 1	100	-	86
Pop 2	94	6	224
Pop 3	99	1	70
Combined	96	4	380

D.12. Do you anticipate that make-up sessions will be required at your school?

Responses on for this item were omitted on thirteen of the observation records.

	% of Responses		
	Yes	No	Total
Pop 1	12	88	84
Pop 2	11	89	220
Pop 3	36	64	67
Combined	16	84	371

D.13. If yes, do you intend to conduct one?

	% of Responses		
	Yes	No	Total
Pop 1	89	11	9
Pop 2	96	4	24
Pop 3	90	10	20
Combined	93	7	53

D.14. Please comment on the selection and training of your Test Administrators. Did it work well? Was the *Test Administrator Manual* useful for training purposes? Were you given time by the Principal to do the training, etc.?

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Generally positive comment	38	91	23	152
b.	No training provided	10	38	7	55
c.	Training very short	4	16	5	25
d.	Straight forward	8	1	2	11
e.	Trained elsewhere	1	5	5	11
f.	The TA manual was useful/clear	-	13	1	14
g.	Would be better to have training in simulated sessions	-	8	-	8
h.	School Coordinator and TA worked together	-	6	1	7
i.	Useful/necessary	1	4	1	6
j.	Manual could be improved	-	2	1	3
k.	Miscellaneous	1	3	2	6
		Total=63	Total=187	Total=48	Total=298

D.15. Please comment on the sampling procedures used to select the students in your school (clarity of instructions, difficulties, time consumed, suggestions for improvement, etc.).

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Not involved with process	17	68	11	96
b.	Straight forward/very clear	21	37	17	75
c.	Complete classes sampled	5	14	2	21
d.	Selected randomly	-	7	3	10
e.	Students not a good representation of the school	-	8	-	8
f.	Time consuming process	1	-	4	5
g.	All advance math students tested in this country (Pop 3)	-	-	3	3
h.	All the classes were taken	-	5	2	7
i.	Not clear	1	1	-	2
j.	Selected on school achievement to give a representative distribution of achievement	-	1	-	1
		Total=45	Total=141	Total=42	Total=228

D.16. Did the students receive any special instructions, motivational talk, or incentives to prepare them for the assessment?

Responses on for this item were omitted by nineteen of the sessions observed.

	% of Responses		
	Yes	No	Total
Pop 1	42	58	83
Pop 2	43	57	212
Pop 3	59	41	70
Combined	46	54	365

D.17. If yes, what special instructions, motivational talk or incentives were provided and by whom?

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Basic information	19	29	15	63
b.	Motivational talk	6	46	13	65
c.	Pamphlet	2	2	2	6
d.	Newsletter	-	1	-	1
e.	Letter to parents	2	3	3	8
f.	Actual incentives	-	1	1	2
		Total=18	Total=82	Total=34	Total=134

D.18. Were students given any opportunity to practice on questions like those in the tests before the testing session?

Responses on for this item were omitted by nine of the sessions observed.

	% of Responses		
	Yes	No	Total
Pop 1	-	100	85
Pop 2	3	97	220
Pop 3	-	100	70
Combined	2	98	375

D.19. If yes, please explain.

		# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	Preliminary training session was conducted on the recommendations of the NRC	2	-	2
b.	Test have been used for assessment previously	-	2	2
c.	"We have orientation tests, or evaluation tests that are very similar"	1	-	1
		Total=3	Total=2	Total=5

D.20. Overall do you feel the TIMSS *School Coordinator Manual* worked well or does it need improvement?

Responses on for this item were omitted by 48 of the sessions observed.

	% of Responses		
	Worked well	Needs to be improved	Total
Pop 1	95	5	79
Pop 2	91	9	195
Pop 3	90	10	62
Combined	92	8	336

D.21. How should the manual be improved?

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Pop 3	# of Comments Combined
a.	The manual should be shorter/more direct	2	12	1	15
b.	Fine as is	1	10	1	12
c.	Check list/flow chart of things to be done	-	2	1	3
d.	The manual left out many things told by trainers	-	1	1	2
e.	More information about TIMSS	2	-	-	2
f.	Should include instructions about leaving the room	-	-	2	2
g.	The manual should provide better explanations of the structure of the sessions	-	2	-	2
h.	More graphic		2	-	2
i.	Instructions about calculators	-	1	-	1
j.	Should include instructions for students who are finished	-	-	1	1
k.	Would have been better if the manual had been translated	1	-	-	1
l.	A remark not to use extra paper for notes and operations	-	1	-	1
		Total=6	Total=31	Total=7	Total=44

D.22. What is your position at this school or school district?

Responses on for this item were omitted by 16 of the sessions observed.

		% of Responses
a.	School: Classroom teacher	24
b.	School: Principal	38
c.	School: Counselor	2
d.	School: Other	21
e.	School: District Staff	4
f.	Other	12
		Total =368

For Questions D.23 through D.28, show the School Coordinator the Class Tracking Form *for the school. Explain that* TIMSS *is interested in having the most accurate sampling information possible.*

D.23. Is this a complete list of the mathematics classes in this grade in this school?

A number of QCMs (79) did not check this item.

	% of Responses		
	Yes	No	Total
Pop 1	97	3	73
Pop 2	91	9	193
Pop 3	97	3	39
Combined	93	7	305

D.24. If no, please explain.

		# of Comments Pop 1	# of Comments Pop 2	# of Comments Combined
a.	There was at least one other class that was not selected	1	8	9
b.	The list is not complete	-	1	1
c.	There was an error in the class tracking form	1	-	1
d.	Miscellaneous	-	4	4
e.	Irrelevant comment	-	4	4
		Total = 2	Total = 13	Total = 19

D.25. To the best of your knowledge, are there any students in this grade level who are *not* in any of these mathematics classes?

A number of QCMs (82) did not check this item.

	% of Responses		
	Yes	No	Total
Pop 1	1	99	73
Pop 2	3	97	191
Pop 3	3	97	38
Combined	2	98	102

D.26. If yes, please explain.

		# of Comments All Pop 2
a.	Four special education students	1
b.	Extra math classes for reinforcement	1
c.	Three students said that they had taken this test earlier the same year	1
		Total = 3

D.27. To the best of your knowledge, are there any students in this grade level in more than one of these mathematics classes?

A number of QCMs (81) did not check this item.

	% of Responses		
	Yes	No	Total
Pop 1	1	99	73
Pop 2	3	97	191
Pop 3	11	89	38
Combined	4	96	302

D.28. If yes, please explain.

Four of the 8 observation records providing comments have comments that indicate the School Coordinator misunderstood the question. For example, one SC interpreted this question to ask "How many math classes in this grade level are in this school?".

	Number of Comments
a. In extra classes to reinforce	3
b. Slow learners need extra classes	1
	Total = 4

D.29. If there were another international assessment, would you be willing to serve as a School Coordinator?

	% of Responses		
	Yes	No	Total
Pop 1	85	15	85
Pop 2	91	9	220
Pop 3	94	6	69
Combined	90	10	374

D.30. That is all the questions I have. On behalf of the Third International Mathematics and Science Study, I want to thank you for your time and effort. Do you have any comments you would like to make before I leave?

202 of the school coordinators provided comments. The comments which are directly related to the study are tabulated below.

Subject of Comments	Nature of Comments	Total
School Coordinator	Appreciative of study, positive experience..........	43
	Glad the study is over, stressful experience.........	2
Teachers	Cooperative, helpful.....................................	1
	Poorly organized, unmotivated and uncooperative, inadequate supervision of students....................	1
Students	Positive about study.....................................	4
	Negative about study.....................................	-
Time	Study was too time consuming........................	6
	Study interfered with school program too much ..	4
	Time allocated to test was inappropriate............	9
	Timing of Study was inappropriate..................	26
Study materials	Excellent, well presented, easy to follow directions, stimulating questions....................................	5
	Questionnaires were too complex or demanding..	14
	Mismatch between test and curriculum..............	45
Results	Make results known......................................	20

APPENDIX G

TIMSS INTERNATIONAL CODING RELIABILITY STUDY
December 4 - 8, 1995
Participant List

Australia
Lynette Beeley
Colin T. Crawford

Bulgaria
Petia Ivanova Assenova
Marionela Zaharieva Simova

Canada
Bill Kokoskin

England
David Harris
Parvin Stewart

France
Josette Le Coq

Germany
Alexander Grass

Hong Kong
Maggie Wong
Chi-Kin Wong

Ireland
Nuala O'Malley
Deirdre Stuart

Lativa
Valdis Apsenieks
Andris Grinfelds

Lithuania
Pranas Gudynas
Gediminas Trakas

New Zealand
Megan Chamberlain
Glenn Chamberlain

Norway
Vegard Brekke
Svein Lie

Philippines
Milagros D. Ibe
Jose A. Fadul

Portugal
Helena Henriques
Filomena de Jesus Ribeiro Cardoso Neres

Romania
Gabriela Noveanu
Viorica Livia Pop

Russian Federation
Svetlana Djukova
Klara Krasnianskaia

Singapore
Lee Ah Huat
Mok Siew Eng

Slovak Republic
Maria Berova
Eva Chmurova

Sweden
Karin Gustafsson
Bjarne Jonsson

Switzerland
Francesca Pedrazzini-Pesce
Sonja Ocshner

United States
John W. Anderson
Jay L. Happel
Connie R. Smith
Lois Yoder
Ina V.S. Mullis
Teresa Smith

APPENDIX H

Contingency Tables and Coding Guides for Items in the International Reliability Study

Spent Fraction of Total Money

Item M1

Mr. Lewis had $360. He spent 7/9 of it. How much money did he have left?

	FREQUENCIES OF MATCHED CODES									
	10	**70**	**71**	**72**	**73**	**75**	**79**	**90**	**99**	**ROW SUM**
10	3525	0	8	0	7	0	22	0	7	3569
70		148	0	0	0	0	27	0	0	175
71			225	0	0	0	34	0	0	259
72				204	0	0	20	0	0	224
73					36	0	13	0	0	49
75						101	10	0	0	111
79							4072	43	10	4125
90								89	76	165
99									473	473

TOTAL VALID COMPARISONS 9150

Coding Guide	
Code	**Response**
Correct Response	
10	80
Incorrect Response	
70	2/9
71	40
72	120
73	180
74	28
75	300
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

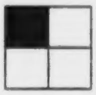
BEST COPY AVAILABLE

Pattern of Boxes: Number of Pieces

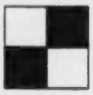
Item M2A

Two boxes of square-shaped cardboard pieces are available to make a larger pattern. There are 4 small squares in each piece.

All pieces in Box 1 look like:



All pieces in Box 2 look like:



In the required pattern, for every piece from Box 2 there are 2 pieces from Box 1.

a) **If 60 pieces from Box 2 are used in the required pattern, how many pieces will be needed altogether?**

b) What fraction of the small squares in the required pattern will be black?

	FREQUENCIES OF MATCHED CODES - M2A								
	10	70	71	72	73	79	90	99	ROW SUM
10	14396	63	0	128	0	34	0	0	14621
70		2137	16	0	0	31	0	0	2184
71			3123	0	0	15	16	0	3154
72				[illegible]	17	82	27	0	13765
73					[illegible]	47	0	0	1554
79						[illegible]	108	22	7710
90							[illegible]	168	312
99								[illegible]	2750

TOTAL VALID COMPARISONS	46050

Codes for Number of Pieces

Coding Guide - M2A	
Code	**Response**
Correct Response	
10	180
Incorrect Response	
70	30
71	90
72	120
73	240
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

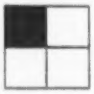
BEST COPY AVAILABLE

Pattern of Boxes: Fraction Black

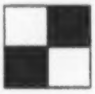
Item M2B

Two boxes of square-shaped cardboard pieces are available to make a larger pattern. There are 4 small squares in each piece.

All pieces in Box 1 look like:



All pieces in Box 2 look like:



In the required pattern, for every piece from Box 2 there are 2 pieces from Box 1.

a) If 60 pieces from Box 2 are used in the required pattern, how many pieces will be needed altogether?

b) What fraction of the small squares in the required pattern will be black?

	FREQUENCIES OF MATCHED CODES - M2B										
	10	19	70	71	72	73	74	79	90	99	ROW SUM
10	5892	123	44	15	72	0	16	156	31	0	6349
19		191	3	0	36	0	167	45	0	0	442
70			1795	4	646	0	15	82	24	60	2626
71				2638	108	44	16	32	0	0	2838
72					10505	0	15	156	9	0	10685
73						1516	15	18	0	0	1549
74							10315	1032	123	39	11509
79								[illegible]	364	34	5194
90									[illegible]	599	1093
99										3765	3765

TOTAL VALID COMPARISONS	**46050**

Codes for Fraction Black

Coding Guide - M2B	
Code	**Response**
Correct Response	
10	1/3
19	A fraction or percent equivalent to 1/3.
Incorrect Response	
70	1/4 OR both 1/4 and 1/2.
71	3/8 or equivalent.
72	1/2 or equivalent.
73	3/4 or equivalent.
74	Any INTEGER.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Chemist Mixes Solution

Item M3

A chemist mixes 3.75 milliliters of solution A with 5.625 milliliters of solution B to form a new solution. How many milliliters does this new solution contain?

	FREQUENCIES OF MATCHED CODES								
	10	19	70	71	72	79	90	99	ROW SUM
10	9366	0	0	43	0	17	0	8	9434
19		0	0	0	0	0	0	0	0
70			568	0	0	8	0	0	576
71				542	0	105	0	0	647
72					685	14	0	0	699
79						1074	20	0	1094
90							6	8	14
99								136	136

TOTAL VALID COMPARISONS	12600

Coding Guide	
Code	**Response**
Correct Response	
10	9.375
19	Other responses equivalent to 9.375
Incorrect Response	
70	8.700 OR 8.7
71	Contains one miscalculated digit.
72	One of the following: 6, 60, 600 OR 6000
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Time To Bake a Cake

Item M4

A cake is put in the oven at 7:20. If the cake takes three quarters of an hour to bake, at what time should it be taken out of the oven?

	FREQUENCIES OF MATCHED CODES										
	10	19	70	71	72	73	74	79	90	99	ROW SUM
10	36008	74	0	21	16	47	0	174	4	0	36344
19		36	0	0	0	0	0	27	2	0	65
70			120	0	0	0	0	0	0	0	120
71				940	0	0	0	45	6	0	991
72					786	0	0	0	0	0	786
73						961	0	0	0	0	961
74							393	28	0	0	421
79								5758	107	15	5880
90									2	15	17
99										465	465

TOTAL VALID COMPARISONS	46050

Coding Guide	
Code	**Response**
Correct Response	
10	8:05
19	Responses equivalent to 8:05.
Incorrect Response	
70	7:50
71	8:00
72	8:10
73	8:15
74	8:35
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Area of a Rectangle

Item M5

The length of a rectangle is 6 cm, and its perimeter is 16 cm. What is the area of the rectangle in square centimeters?

	FREQUENCIES OF MATCHED CODES									
	10	70	71	72	73	74	79	90	99	ROW SUM
10	19485	16	50	0	16	22	203	1	0	19793
70		3310	15	14	32	0	112	0	0	3483
71			2442	0	0	0	0	0	0	2442
72				648	0	0	1	0	0	649
73					1554	0	15	0	16	1585
74						5489	328	5	0	5822
79							9639	63	114	9816
90								127	186	313
99									[illegible]	2082

TOTAL VALID COMPARISONS	45985

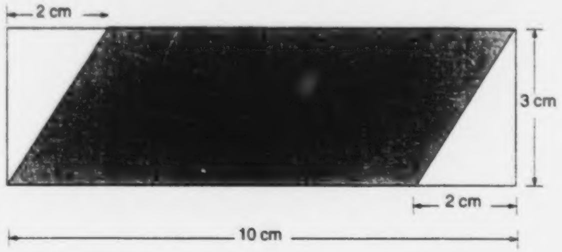
Coding Guide	
Code	**Response**
Correct Response	
10	12
Incorrect Response	
70	22
71	24
72	48
73	60
74	96 or indication of 6x16
79	Other incorrect
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Area of Parallelogram

Item M6

The figure show a shaded parallelogram inside a rectangle.



What is the area of the parallelogram?

	FREQUENCIES OF MATCHED CODES									
	10	70	71	72	73	74	79	90	99	ROW SUM
10	5313	9	3	0	1	0	73	0	0	5399
70		185	0	0	1	0	36	0	0	222
71			1217	0	0	0	59	0	0	1276
72				277	0	0	46	0	0	323
73					685	0	20	0	0	705
74						137	22	0	0	159
79							4154	64	28	4246
90								[illegible]	6	19
99									251	251

TOTAL VALID COMPARISONS	12600

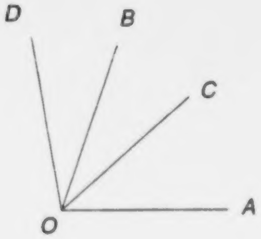
Coding Guide	
Code	**Response**
Correct Response	
10	24
Incorrect Response	
70	10
71	18
72	26
73	30
74	60
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Degrees of Angle

Item M7

In the figure, the measure of g*AOB* is 70°, the measure of g*COD* is 60°, and the measure of g*AOD* is 100°.



What is the measure of g*COB?*

	FREQUENCIES OF MATCHED CODES										
	10	70	71	72	73	74	75	79	90	99	ROW SUM
10	5396	0	0	0	0	16	0	61	11	0	5484
70		100	0	0	0	0	0	8	0	0	108
71			732	0	0	0	0	96	0	0	828
72				1092	0	23	0	9	0	8	1132
73					368	7	0	65	0	0	440
74						818	8	22	0	8	856
75							390	14	6	8	418
79								2451	122	29	2602
90									166	145	311
99										421	421

TOTAL VALID COMPARISONS	12600

Coding Guide	
Code	**Response**
Correct Response	
10	30
Incorrect Response	
70	20
71	35
72	40
73	45
74	50
75	60 OR 70
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Decimal as a Fraction in Lowest Terms

Item M8

Write 0.28 as a fraction reduced to its lowest terms.

	FREQUENCIES OF MATCHED CODES									
	10	70	71	72	73	74	79	90	99	ROW SUM
10	5813	45	1	0	0	0	66	6	0	5931
70		607	43	0	0	32	68	9	5	764
71			138	15	0	17	69	3	0	242
72				318	0	39	69	1	0	427
73					926	0	60	0	0	986
74						87	248	2	0	337
79							2337	45	21	2403
90								167	280	447
99									1063	1063

TOTAL VALID COMPARISONS	12600

Coding Guide	
Code	**Response**
Correct Response	
10	7/25
Incorrect Response	
70	28/100 OR 14/50
71	Any fractions other than 28/100 with 28 as numerator.
72	Any fractions with 28 as denominator.
73	2/8 OR 1/4
74	Any expression which mixes decimal notation into the fraction.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Amount Sue Paid

Item M9

Peter bought 70 items and Sue bought 90 items. Each item cost the same and the items cost $800 altogether. How much did Sue pay?

	FREQUENCIES OF MATCHED CODES									
	10	70	71	72	73	74	79	90	99	ROW SUM
10	3143	0	7	0	14	21	62	2	0	3249
70		26	0	0	15	0	38	0	0	79
71			732	0	0	0	39	0	0	771
72				183	0	0	8	0	0	191
73					539	0	65	0	7	611
74						414	72	10	0	496
79							2881	64	16	2961
90								122	109	231
99									561	561

TOTAL VALID COMPARISONS	9150

Coding Guide	
Code	**Response**
Correct Response	
10	450
Incorrect Response	
70	5
71	400
72	420
73	500
74	600
79	Other incorrect
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Recording Songs: Estimate of Time

Item M10A

Teresa wants to record 5 songs on tape. The length of time each song plays for is shown in the table.

Song	Amount of Time
1	2 minutes 41 seconds
2	3 minutes 10 seconds
3	2 minutes 51 seconds
4	3 minutes
5	3 minutes 32 seconds

a) **ESTIMATE to the nearest minute the total time taken for all five songs to play.**
b) Explain how this estimate was made.

	FREQUENCIES OF MATCHED CODES - M10A									
	10	11	70	71	72	73	79	90	99	ROW SUM
10	16532	407	7	5	114	0	425	14	0	17504
11		9518	0	71	13	0	94	32	0	9728
70			1042	0	7	0	214	1	0	1264
71				3044	13	14	276	2	0	3349
72					1435	0	508	0	0	1943
73						501	44	0	0	545
79							9848	245	16	10109
90								12	0	12
99									1484	1484

TOTAL VALID COMPARISONS	45938

Codes for Total Estimate

Coding Guide - M10A	
Code	**Response**
Correct Response	
10	15 minutes.
11	16 minutes.
Incorrect Response	
70	13 minutes.
71	14 minutes.
72	15 min. 14 sec.
73	17 minutes.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Recording Songs: Explanation of Estimate

Item M10B

Teresa wants to record 5 songs on tape. The length of time each song plays for is shown in the table.

Song	Amount of Time
1	2 minutes 41 seconds
2	3 minutes 10 seconds
3	2 minutes 51 seconds
4	3 minutes
5	3 minutes 32 seconds

a) ESTIMATE to the nearest minute the total time taken for all five songs to play.
b) Explain how this estimate was made.

	FREQUENCIES OF MATCHED CODES - M10B										
	10	**11**	**12**	**13**	**19**	**70**	**71**	**79**	**90**	**99**	**ROW SUM**
10	2624	44	2549	85	360	87	39	198	56	124	6166
11		49	342	22	174	19	2	185	4	0	797
12			8873	870	2057	280	188	3487	388	59	16202
13				1805	677	47	166	722	31	27	3475
19					788	79	63	1207	72	15	2224
70						63	27	531	11	36	668
71							718	960	3	37	1718
79								9487	1063	198	10748
90									140	217	357
99										3695	3695

TOTAL VALID COMPARISONS	**46050**

Codes for Explanation

Coding Guide - M10B	
Code	**Response**
Correct Response	
10	Each amount of time is correctly rounded to whole minutes before adding.
11	Each amount of time is correctly rounded to nearest 5,10,15 or 30 seconds.
12	No calculation shown. Statements may include "rounded off to nearest minute", "rounded the numbers up and down" or similar expressions.
13	Adds correctly and then rounds off from 15 min. 14 sec.
19	Other correct.
Incorrect Response	
70	Each amount of time is rounded off, but one or more rounding is incorrect.
71	Rounds off from 14 min. 34 sec.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Apples in Box: Answers

Item M11A

a) There are 54 kilograms of apples in two boxes. The second box of apples weighs 12 kilograms more than the first. How many kilograms of apples are in each box?

b) Show vour work.

	FREQUENCIES OF MATCHED CODES - M11A									
	20	**10**	**11**	**70**	**71**	**72**	**79**	**90**	**99**	**ROW SUM**
20	4668	7	17	16	16	0	8	0	8	4740
10		13	49	0	0	0	30	0	0	92
11			216	0	12	0	206	8	12	454
70				1960	0	0	104	5	3	2072
71					483	12	185	25	8	713
72						16	14	0	6	36
79							3386	214	155	3755
90								55	38	93
99									637	637

TOTAL VALID COMPARISONS	**12592**

Codes for Correctness

Coding Guide - M11A	
Code	**Response**
Correct Response	
20	33 kg AND 21 kg.
Partial Response	
10	Follows the right steps but makes a small arithmetic error resulting in an incorrect answer.
11	Either 33 kg OR 21 kg, with or without another incorrect weight.
Incorrect Response	
70	15 kg AND 39 kg.
71	One of the answers is 42 kg.
72	15 kg AND 27 kg.
79	Other incorrect
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Apples in Box: Method

Item M11B

a) There are 54 kilograms of apples in two boxes. The second box of apples weighs 12 kilograms more than the first. How many kilograms of apples are in each box?

b) **Show your work.**

	FREQUENCIES OF MATCHED CODES - M11B									
	10	11	12	19	70	71	79	90	99	ROW SUM
10	1148	23	78	30	0	61	8	8	0	1356
11		722	112	40	6	171	38	21	1	1111
12			2198	35	0	125	8	6	0	2372
19				72	47	32	20	7	5	183
70					108	56	81	28	86	359
71						3117	1748	113	20	4998
79							1286	120	69	1475
90								46	37	83
99									663	663

TOTAL VALID COMPARISONS	12600

Codes for Method

Coding Guide - M11B	
Code	**Response**
Correct Response	
10	An equation with an unknown variable explicitly shown.
11	Method: divide 54 by 2, then add 6 to 27 to get 33 and subtract 6 from 27 to get 21. [Addition and subtraction of 6 need not be shown if student has arrived at the correct solution].
12	Method: subtract 12 from 54 to obtain 42, then divide by 2 to obtain 21kg and then add 12 to get 33 kg. [Addition of 12 to obtain 33 need not be shown if student arrived at the correct solution].
19	Other fully satisfactory solution including "guess and check" with justification that 21 + 33 = 54.
Incorrect Response	
70	No method is shown.
71	Method shown is inadequate, but begins in appropriate manner.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Rounded versus Actual Weight

Item M12

Rounded to the nearest 10 kg the weight of a dolphin was reported as 170 kg. Write down a weight that might have been the actual weight of the dolphin.

	FREQUENCIES OF MATCHED CODES											
	10	**11**	**12**	**13**	**70**	**71**	**72**	**73**	**79**	**90**	**99**	**ROW SUM**
10	6450	8	24	86	50	8	8	5	207	0	0	6846
11		473	15	0	0	0	0	15	0	0	0	503
12			1704	8	18	0	16	7	8	0	0	1761
13				208	0	0	0	0	26	0	0	234
70					51	0	15	4	91	0	0	161
71						801	24	19	43	0	0	887
72							493	11	12	0	0	516
73								18	134	1	0	153
79									1280	7	0	1287
90										74	28	102
99											150	150

TOTAL VALID COMPARISONS	12600

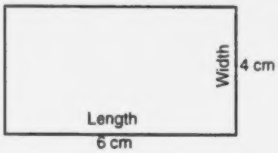
Coding Guide

Code	Response
Correct Response	
10	Number within the interval 165<X<170.
11	170
12	Number within the interval 170<X<175.
13	Two or more numbers within the interval 165<X<175.
Incorrect Response	
70	Number within the interval 175<X<180.
71	150 OR 200
72	160 OR 180
73	Result of converting 170 kg to other units.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Draw New Rectangle

Item M13A



a. **In the space below, draw a new rectangle whose length is one and one half times the length of the rectangle above, and whose width is half the width of the rectangle above. Show the length and width of the new rectangle in centimeters on the figure.**

b. What is the ratio of the area of the new rectangle to the area of the first one? Show your work.

	FREQUENCIES OF MATCHED CODES - M13A											
	20	10	11	70	71	72	73	74	79	90	99	ROW SUM
20	3819	82	510	0	0	0	4	12	10	0	0	4437
10		42	18	0	0	0	7	0	0	1	0	68
11			638	4	0	0	25	4	8	3	0	682
70				274	0	0	253	0	55	4	0	586
71					162	0	45	0	15	0	0	222
72						834	52	0	56	0	0	942
73							1162	16	220	38	8	1444
74								846	130	10	8	994
79									2354	90	28	2472
90										64	126	190
99											563	563

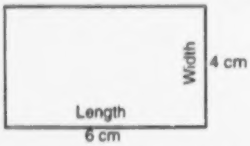
TOTAL VALID COMPARISONS	12600

Codes for Drawing

Coding Guide - M13A	
Code	**Response**
Correct Response	
20	9 cm and 2 cm. Correct drawing shown.
Partial Response	
10	9 cm and 2 cm. Drawing is incorrect or missing.
11	The length and/or width is not given or is incorrect. Correct drawing is shown.
Incorrect Response	
70	15 cm and 2 cm. Explicitly written or implicit from the drawing.
71	7.5 cm and 2 cm. Explicitly written or implicit from the drawing.
72	3 cm and 2 cm. Explicitly written or implicit from the drawing.
73	2 cm width and a length equal to any other numbers except those given above. Explicitly written or implicit from the drawing.
74	9 cm length and a width equal to any other numbers than those given above. Explicitly written or implicit from the drawing.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Ratio of Rectangle Areas

Item M13B



a) In the space below, draw a new rectangle whose length is one and one half times the length of the rectangle above, and whose width is half the width of the rectangle above. Show the length and width of the new rectangle in centimeters on the figure.

b) What is the ratio of the area of the new rectangle to the area of the first one? Show your work.

	FREQUENCIES OF MATCHED CODES - M13B												
	20	21	10	11	12	13	14	19	70	79	90	99	ROW SUM
20	1527	18	136	134	6	4	1	10	7	53	1	0	1897
21		316	18	31	0	281	40	5	25	189	19	5	929
10			592	87	9	13	10	0	12	29	6	0	758
11				230	46	10	16	13	3	39	1	0	358
12					128	0	7	0	0	4	0	0	139
13						194	52	6	24	234	4	2	516
14							33	0	20	65	2	0	120
19								0	1	13	0	0	14
70									232	904	45	31	1212
79										3074	488	290	3852
90											138	222	360
99												2437	2437

TOTAL VALID COMPARISONS	12592

Codes for Ratio and Areas

Coding Guide - M13B	
Code	**Response**
Correct Response	
20	3:4, 3/4 or equivalent. The areas are 18 cm2 and 24 cm2.
21	The ratio is NOT 3:4 but areas and ratio of part (b) are consistent with response in part (a).
Partial Response	
10	4:3 or equivalent. (Ratio is reversed.) The areas are 18 cm2 and 24 cm2.
11	An incorrect ratio or no ratio is given. The areas are 18 cm2 and 24 cm2.
12	The difference between the areas, 6, is given instead of the ratio. The areas are 18 cm2 and 24 cm2.
13	Areas are NOT 18 cm2 and 24 cm2 but are consistent with response in part a) and an incorrect ratio or no ratio is given.
14	Areas are NOT 24 cm2 and 18 cm2 but are consistent with response in part a) and a difference consistent with those areas is given instead of the ratio.
Incorrect Response	
70	Focuses exclusively on the ratios of lengths and widths between the given rectangle and No areas are shown.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Cheaper Office Rental

Item M14

The following two advertisements appeared in a newspaper in a country where the units of currency are *zeds* .

BUILDING A
Office space available

85-95 square meters
475 *zeds* per month

100 - 120 square meters
800 *zeds* per month

BUILDING B
Office space available

35-260 square meters
90 *zeds* per square meters
per year

If a company is interested in renting an office of 110 square meters in that country for a year, at which office building, A or B, should they rent the office in order to get the lower price? Show your work.

	FREQUENCIES OF MATCHED CODES														
	30	**39**	**20**	**21**	**10**	**11**	**12**	**16**	**19**	**70**	**71**	**79**	**90**	**99**	**ROW SUM**
30	2351	223	94	58	17	0	14	0	6	0	6	0	0	0	2769
39		0	36	3	16	0	2	0	0	0	0	0	0	0	57
20			794	23	192	21	98	8	3	3	0	2	4	0	1148
21				211	3	0	42	0	15	17	1	0	0	0	289
10					991	108	4	118	37	13	4	57	22	3	1357
11						398	0	42	27	0	0	2	23	69	561
12							812	0	0	223	19	34	20	0	1108
16								45	16	4	4	6	0	0	75
19									1	1	1	0	3	1	7
70										1145	382	201	18	0	1746
71											1134	136	37	126	1433
79												151	84	16	251
90													153	179	332
99														1467	1467

TOTAL VALID COMPARISONS 12600

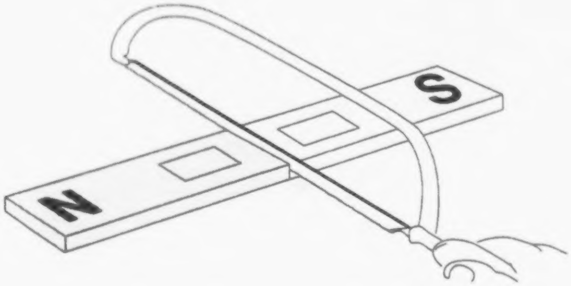
Coding Guide	
Code	**Response**
Correct Response	
30	Building A. Correct calculation of rents for both buildings. 9600/800 AND 9900/825, OR 825 to compare with the 800 given.
39	Other correct.
Partial Response	
20	Building A. Correct calculation of rent for Building A OR B but not both.
21	Building B OR building is not named. Correct calculation of rents for both buildings.
Minimal Response	
10	Building A. Calculations or explanation are incorrect or inadequate.
11	Building A. No work shown.
12	Building B, OR building is not named. Correct calculation of rent for Building A OR B but not both.
16	Building A. Explanation is given only in the form of extracts from the advertisements.
19	Other minimal.
Incorrect Response	
70	Building B. Incorrect or inadequate calculations.
71	Building B. No work shown.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Magnets

Item S1

A bar magnet is cut in two with a hacksaw. Write an "N" or an "S" in each box on the diagram to show the polarity of the cut ends.



	FREQUENCIES OF MATCHED CODES								
	10	19	70	71	72	79	90	99	ROW SUM
10	5466	0	42	0	0	0	0	20	5528
19		0	0	0	0	0	0	0	0
70			2459	24	0	0	0	8	2491
71				57	0	0	0	0	57
72					21	0	0	0	21
79						181	22	15	218
90							19	36	55
99								708	708

TOTAL VALID COMPARISONS	9078

Coding Guide	
Code	**Response**
Correct Response	
10	S - N are written in the open boxes in this order.
19	Other correct.
	Incorrect Response
70	N -S are written in the open boxes in this order.
71	N - N are written in the open boxes.
72	S - S are written in the open boxes.
79	Other incorrect: Including a single N or S.
	Nonresponse
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Jose's Influenza

Item S2

José caught influenza. Write down one way he could have caught it.

	FREQUENCIES OF MATCHED CODES								
	10	11	12	19	70	79	90	99	ROW SUM
10	5183	2258	500	576	198	211	6	13	8945
11		6400	3275	398	114	623	8	0	10818
12			7984	160	190	624	30	55	9043
19				170	89	783	20	1	1063
70					7677	930	58	80	8745
79						3259	328	68	3655
90							295	462	757
99								3009	3009

TOTAL VALID COMPARISONS	46035

Coding Guide	
Code	**Response**
Correct Response	
10	Refers explicitly to transmission of germs.
11	Refers implicitly to transmission of germs by sneezing/coughing or close contact.
12	States only that he got it from someone (who had the flu).
19	Other correct.
Incorrect Response	
70	Refers to being too cold.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

246

BEST COPY AVAILABLE

Melting Ice Cubes

Item S3

A glass of water with ice cubes in it has a mass of 300 grams. What will the mass be immediately after the ice has melted? Explain your answer.

	FREQUENCIES OF MATCHED CODES										
	20	10	11	70	71	72	73	79	90	99	ROW SUM
20	2824	431	34	76	0	10	0	1	0	5	3381
10		154	28	18	0	19	0	24	2	1	246
11			82	7	0	0	0	0	0	0	89
70				1847	66	74	2	121	9	0	2119
71					72	0	0	29	0	0	101
72						1212	49	102	7	0	1370
73							73	25	8	0	106
79								331	107	0	438
90									123	112	235
99										1065	1065

TOTAL VALID COMPARISONS	9150

Coding Guide	
Code	**Response**
Correct Response	
20	300 g with a good explanation.
Partial Response	
10	300 g. Explanation is inadequate.
11	300 g. No explanation.
Incorrect Response	
70	More than 300 grams with explanation.
71	More than 300 g. No explanation.
72	Less than 300 g. With explanation.
73	Less than 300 g. No explanation.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

How Computers Help

Item S4

Write down one example of how computers help people do their work.

	FREQUENCIES OF MATCHED CODES											
	10	**11**	**12**	**13**	**14**	**19**	**70**	**71**	**76**	**79**	**90**	**99**	**ROW SUM**
10	1622	25	128	67	277	712	4	62	2	130	12	1	3042
11		520	18	59	109	190	0	20	0	40	4	7	967
12			3716	149	583	433	0	88	5	69	0	7	5050
13				111	91	81	1	19	3	26	2	0	334
14					664	290	1	7	0	17	1	0	980
19						1288	18	134	15	170	18	3	1646
70							1	2	0	3	1	0	7
71								52	22	49	4	1	128
76									3	3	0	0	6
79										41	7	4	52
90											0	23	23
99												365	365

TOTAL VALID COMPARISONS	**12600**

Coding Guide	
Code	**Response**
Correct Response	
10	Refers to writing OR editing text.
11	Refers to doing calculations OR doing them faster.
12	Refers to computer storing or retrieving information (promptly).
13	Refers to using computers for instruction.
14	Response refers to any combination of two or more of codes 10-13.
19	Other correct:
Incorrect Response	
70	Playing games such as Nintendo.
71	Vague references to "everything" or some similar expression.
76	Merely repeats information in the stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Life on Another Planet

Item S5

Jane and Mario were discussing what it might be like to live on other planets. Their science teacher gave them data about the earth and an imaginary planet, Athena. The table shows these data.

	Earth	Athena
Atmospheric Conditions	21% oxygen 0.03% carbon dioxide 78% nitrogen ozone layer	10% oxygen 80% carbon dioxide 5% nitrogen no ozone layer
Distance from a star Like the Sun	148,640,000 km	103,600,000 km
Rotation on Axis	1 day	200 days
Revolution Around Sun	365 1/4 days	200 days

Write down one important reason why it would be difficult for humans to live on Athena if it existed.

	FREQUENCIES OF MATCHED CODES										
	10	**11**	**12**	**13**	**14**	**19**	**70**	**79**	**90**	**99**	**ROW SUM**
10	883	134	61	1	456	131	22	333	19	0	2040
11		9092	19	251	2214	298	13	596	37	0	12520
12			159	37	113	74	42	260	11	0	696
13				7506	1529	130	77	793	231	0	10266
14					[illegible]	312	52	782	40	0	16676
19						184	162	381	12	0	739
70							55	102	1	0	158
79								1246	129	22	1397
90									67	24	91
99										1467	1467

TOTAL VALID COMPARISONS	**46050**

Coding Guide	
Code	**Response**
Correct Response	
10	States that there would be too much carbon dioxide.
11	States that there would be too little oxygen to breathe.
12	Refers to bound rotation, that is, the periods of revolution around the planet's own axis and rotation around its sun are the same. Hence, one side of the planet is always facing the sun and therefore is hot while the other side is always dark and cold.
13	States that there is no ozone.
14	Any combination of above codes, 10-13.
19	Other correct.
Incorrect Response	
70	States that it is too close to a star, without further explanation.
79	Other incorrect or seriously incomplete.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

How Air Exists

Item S6

Air is colorless, odorless, and tasteless. Describe one way that air can be shown to exist.

	FREQUENCIES OF MATCHED CODES													
	10	11	12	13	14	15	19	70	72	76	79	90	99	ROW SUM
10	11119	337	28	204	151	90	533	538	153	42	977	134	0	14306
11		405	0	0	0	2	27	19	2	0	55	5	0	515
12			1043	240	12	14	193	32	34	0	51	0	16	1635
13				[illegible]	99	25	225	172	45	0	133	18	0	4283
14					[illegible]	43	185	32	0	0	184	35	0	825
15						867	244	68	111	0	176	3	0	1469
19							204	664	128	4	513	23	0	1536
70								8893	497	3	833	26	17	10269
72									2524	0	271	9	0	2804
76										14	116	22	6	158
79											2531	325	51	2907
90												148	265	413
99													4930	4930

TOTAL VALID COMPARISONS	46050

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that you can feel or see effects of air movement.
11	Mentions that (light) things fall slowly.
12	Refers to the fact that air can be weighed.
13	Mentions that balloons or tires, etc. can be filled with air.
14	Refers to air pressure.
15	Refers to being able to 'see' air.
19	Other correct.
Incorrect Response	
70	We can breathe air.
71	Refers only to the need of oxygen or air for life and other processes.
72	Refers to seeing water vapor.
76	Merely repeats information in the stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

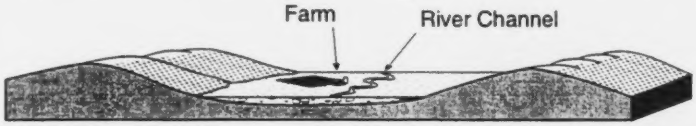
BEST COPY AVAILABLE

River on Plain: Good Place

Item S7A

The diagram shows a river flowing through a wide plain. The plain is covered with several layers of soil and sediment.

a. **Write down one reason why this plain is a good place for farming.**
b. Write down one reason why this plain is NOT a good place for farming.



	Frequencies of Matched Codes - S7A									
	10	11	12	19	70	76	79	90	99	ROW SUM
10	2437	758	147	219	4	243	113	37	0	3958
11		2723	118	256	8	101	104	3	0	3313
12			805	112	12	9	87	7	0	1032
19				120	17	0	167	8	0	312
70					0	2	31	2	0	35
76						43	23	19	0	85
79							130	30	2	162
90								11	10	21
99									232	232

TOTAL VALID COMPARISONS	9150

Codes for Good Place

Coding Guide - S7A	
Code	**Response**
Correct Response	
10	Mentions that the soil is fertile (good), abundant.
11	Mentions that there is a river (for irrigation, water for animals).
12	Mentions that there is plenty of space or flat areas for farm land.
19	Other correct:
Incorrect Response	
70	Does not address the issue of farming.
76	Merely repeats information in stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

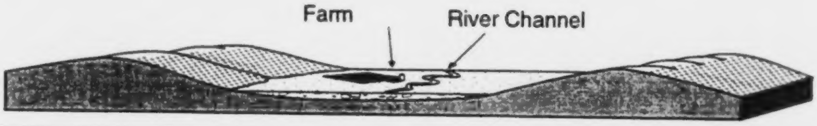
River on Plain: Bad Place

Item S7B

The diagram shows a river flowing through a wide plain. The plain is covered with several layers of soil and sediment.

a. Write down one reason why this plain is a good place for farming.

b. Write down one reason why this plain is NOT a good place for farming.



	Frequencies of Matched Codes - S7B											
	10	11	19	70	71	72	73	76	79	90	99	ROW SUM
10	4294	154	160	6	32	70	11	22	199	11	0	4959
11		447	23	0	2	38	3	0	37	5	0	555
19			201	49	18	122	39	7	280	15	0	731
70				58	27	6	0	2	55	4	0	152
71					[illegible]	25	0	14	115	4	0	189
72						[illegible]	14	20	97	3	0	408
73							[illegible]	74	138	5	0	789
76								22	83	10	0	115
79									[illegible]	83	2	725
90										[illegible]	32	86
99											[illegible]	441

TOTAL VALID COMPARISONS	9150

Codes for Bad Place

Coding Guide - S7B	
Code	**Response**
Correct Response	
10	Mentions the possibility of flooding, or that the soil will be too wet.
11	Mentions the possibility of wind or water erosion.
19	Other correct:
Incorrect Response	
70	Mentions that it is an undesirable place to live: boring/lonesome/ugly.
71	Does not address the issue of farming.
72	Refers to problems due to surrounding mountains.
73	Refers to sediment, soil, being rocky and negative.
76	Merely repeats information in stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Ozone Layer

Item S8

Write down one reason why the ozone layer is important for all living things on Earth.

	FREQUENCIES OF MATCHED CODES													
	10	11	12	19	70	71	72	73	74	76	79	90	99	ROW SUM
10	11590	254	182	88	47	81	59	45	14	0	104	0	0	12464
11		6135	1044	376	201	1320	842	24	70	0	425	24	0	10461
12			1820	80	35	697	254	34	3	0	148	18	0	3089
19				14	24	185	197	2	2	1	74	7	0	506
70					296	561	127	8	50	2	379	24	0	1447
71						2564	703	64	81	4	442	12	0	3870
72							2322	360	195	58	1321	38	0	4294
73								3015	616	2	726	9	0	4368
74									452	4	564	11	0	1031
76										4	37	41	3	85
79											2020	192	23	2235
90												227	78	305
99													1775	1775

TOTAL VALID COMPARISONS	45930

Coding Guide	
Code	**Response**
Correct Response	
10	Refers to protection against the UV radiation from the sun.
11	Refers to protection against dangerous or too strong radiation from the sun but does not mention UV.
12	Mentions that the ozone layer protects humans so we do not get sunburned/skin cancer. NOTE: If UV is mentioned, code 10.
19	Other correct.
Incorrect Response	
70	Confuses the effect of the ozone layer with the greenhouse effect.
71	Confuses protection against heat.
72	Refers only vaguely to protection.
73	Refers to or confuses oxygen, O2 with ozone, O3.
74	Sees the ozone layer as a barrier for the atmosphere.
76	Merely repeats information in the stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Digestion in the Mouth

Item S9

What digestive substance is found in the mouth? What does it do?

	FREQUENCIES OF MATCHED CODES												
	20	21	22	29	10	11	19	70	71	79	90	99	ROW SUM
20	8334	3025	137	625	1971	1	218	0	122	53	48	8	14542
21		5871	393	299	980	6	77	31	82	22	12	0	7773
22			512	57	32	38	12	0	3	2	0	0	656
29				81	738	6	130	8	31	101	1	16	1112
10					5576	1	427	41	134	156	83	33	6451
11						0	3	0	1	8	0	0	12
19							205	4	69	243	12	0	533
70								692	15	20	0	0	727
71									6607	921	22	15	7565
79										1214	77	12	1303
90											178	202	380
99												4996	4996

TOTAL VALID COMPARISONS	46050

Coding Guide

Code	Response
Correct Response	
20	Names saliva and explains that it makes the food moist or soft [Mechanical process].
21	Names saliva and explains that it breaks down the starch or food. [Chemical process].
22	Names enzymes and explains that they break down the starch or food. [Chemical process].
29	Other correct: Names a substance and provides a reasonable explanation.
Partial Response	
10	Names saliva but with no description or with an incorrect description of what it does.
11	Names enzymes but with no description or with an incorrect description of what they do, such as it digests starch.
19	Other partially correct.
Incorrect Response	
70	Acid. With or without description.
71	Teeth, tongue, etc. With or without description.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Size of Sun and Moon

Item S10

The Sun is bigger than the Moon, but they appear to be about the same size when you look at them from the Earth. Why is this?

	FREQUENCIES OF MATCHED CODES											
	20	**29**	**10**	**11**	**19**	**70**	**71**	**76**	**79**	**90**	**99**	**ROW SUM**
20	5402	36	227	0	85	3	29	0	57	5	10	5854
29		0	16	0	4	0	0	0	2	0	0	22
10			889	10	125	26	4	0	527	8	0	1589
11				40	0	1	2	0	22	0	0	65
19					12	6	8	0	76	0	0	102
70						401	16	5	109	1	0	532
71							131	0	64	0	0	195
76								3	24	0	0	27
79									551	20	7	578
90										10	0	10
99											176	176

TOTAL VALID COMPARISONS	9150

Coding Guide	
Code	**Response**
Correct Response	
20	Mentions that the sun is farther away than the moon. Comparative language is used.
29	Other correct comparing apparent sizes.
Partial Response	
10	Refers to distance, but response is general, not specified.
11	Refers to the sun being higher up than the moon.
19	Other acceptable but incomplete or slightly erroneous responses.
Incorrect Response	
70	Response includes some reference to the light.
71	States that the sun is closer than the moon.
76	Merely repeats the information in the stem.
79	Other incorrect:
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Carbon Dioxide Fire Extinguishers

Item S11

Carbon dioxide is the active material in some fire extinguishers. How does carbon dioxide extinguish a fire?

	FREQUENCIES OF MATCHED CODES								
	10	**11**	**19**	**70**	**71**	**79**	**90**	**99**	**ROW SUM**
10	3508	297	340	2	42	542	18	0	4749
11		490	116	29	59	365	1	0	1060
19			43	3	16	165	4	0	231
70				296	10	126	3	0	435
71					265	286	10	0	561
79						1086	89	0	1175
90							111	36	147
99								764	764

TOTAL VALID COMPARISONS	9122

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that carbon dioxide keeps oxygen away; response includes explicit reference to oxygen.
11	Mentions that carbon dioxide keeps "air" away.
19	Other correct.
Incorrect Response	
70	Mentions that carbon dioxide cools down the fire.
71	Refers to a material in carbon dioxide.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Thirsty on a Hot Day

Item S12

Write down the reason why we get thirsty on a hot day and have to drink a lot.

	FREQUENCIES OF MATCHED CODES												
	10	11	12	13	19	70	71	72	76	79	90	99	ROW SUM
10	1273	632	519	223	95	97	25	10	2	31	13	0	2920
11		11322	119	5761	644	199	1255	299	44	242	4	5	19894
12			281	136	94	292	32	11	2	34	1	0	883
13				4129	500	277	1661	327	74	467	39	62	7536
19					111	229	556	48	15	155	3	2	1119
70						2210	567	76	178	525	49	27	3632
71							4244	139	127	763	26	14	5313
72								2030	27	418	12	0	2487
76									181	266	8	0	455
79										1050	103	17	1170
90											2	31	33
99												488	488

TOTAL VALID COMPARISONS	45930

Coding Guide	
Code	**Response**
Correct Response	
10	Refers to perspiration and its cooling effect and the need to replace lost water.
11	Refers to perspiration and only replacement of lost water.
12	Refers to perspiration and only its cooling effect.
13	Refers to perspiration only.
19	Other acceptable explanation.
Incorrect Response	
70	Refers to body temperature (being too hot) but does not answer why we get thirsty.
71	Refers only to drying of the body.
72	Refers to getting more energy by drinking more water.
76	Merely repeats the information in the stem.
79	Other incorrect:
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Advantage of Two Eyes

Item S13

What is the advantage of having two eyes to see with rather than one?

	FREQUENCIES OF MATCHED CODES									
	10	11	12	19	70	71	79	90	99	ROW SUM
10	1827	1410	41	729	83	45	854	33	17	5039
11		23115	1057	2987	2041	186	2755	56	0	32197
12			1340	387	35	58	60	0	0	1880
19				582	352	54	1579	26	0	2593
70					643	19	492	12	0	1166
71						135	115	0	0	250
79							1553	60	0	1613
90								10	0	0
99									1296	1296

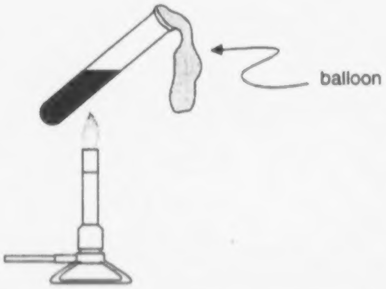
TOTAL VALID COMPARISONS	46034

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that two eyes allow depth perception or better perception of distance.
11	Mentions that two eyes allow seeing more or a wider field of vision.
12	Mentions that with two eyes one is still working if one eye is damaged.
19	Other correct.
Incorrect Response	
70	Mentions seeing twice as much.
71	Refers to energy or effort.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Heating Tube with Balloon Over It

Item S14

The water in a tube is heated, as shown in the diagram. As the water is heated, the balloon increases in size. Explain why.



	FREQUENCIES OF MATCHED CODES									
	10	11	19	70	71	76	79	90	99	ROW SUM
10	156	296	176	48	37	0	161	2	0	876
11		2676	469	189	58	0	488	6	3	3889
19			134	226	170	15	429	6	1	981
70				339	65	13	327	7	0	751
71					112	12	221	2	0	347
76						37	180	7	2	226
79							1608	91	5	1704
90								68	34	102
99									274	274

TOTAL VALID COMPARISONS	9150

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions explicitly that expansion is due to increased pressure of air/gas/water vapor when tube is heated.
11	States that the water evaporates.
19	Other correct.
Incorrect Response	
70	Mentions that hot air [or gas] always rises.
71	Mentions that air particles [or molecules] expand when heated.
76	Merely repeats information in the stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

Diagram: Rain from Another Place

Item S15

Draw a diagram to show how the water that falls as rain in one place may come from another place that is far away.

	FREQUENCIES OF MATCHED CODES									
	20	**10**	**11**	**12**	**19**	**70**	**79**	**90**	**99**	**ROW SUM**
20	16095	1046	2238	370	553	461	288	71	18	21140
10		3355	175	90	296	3082	770	105	1	7874
11			777	16	159	277	235	32	0	1496
12				102	26	30	48	5	0	211
19					28	273	225	20	0	546
70						6439	1925	441	14	8819
79							1293	240	1	1534
90								322	252	574
99									3856	3856

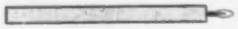
TOTAL VALID COMPARISONS	**46050**

Coding Guide	
Code	**Response**
Correct Response	
20	Response includes the three following steps: i. Evaporation of water from a source. ii. Transportation of water as vapor/clouds to another place. iii. Precipitation in other places.
Partial Response	
10	As in code 20 but response does not mention evaporation.
11	As in code 20, but response does not mention transportation.
12	As in code 20, but response does not mention precipitation.
19	Other partially correct.
Incorrect Response	
70	Response indicates precipitation only; it may use vertical or diagonal lines.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

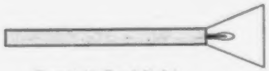
Flashlight with Reflector

Item S16

Jim and Sandy each make a flashlight from identical batteries and bulbs. Sandy's flashlight contains a reflector, while Jim's does not.



Jim's flashlight



Sandy's flashlight

Which flashlight shines more light on a wall 5 meters away, Jim's or Sandy's? Explain your answer.

	FREQUENCIES OF MATCHED CODES												
	10	11	12	19	70	71	72	73	76	79	90	99	ROW SUM
10	1489	68	360	1014	13	16	10	6	256	1132	19	0	4383
11		53	221	44	4	6	7	0	2	53	1	0	391
12			981	118	4	25	15	0	16	280	7	0	1446
19				347	6	5	3	7	196	1105	16	0	1685
70					52	54	102	0	9	38	2	0	257
71						152	202	8	1	71	4	0	438
72							162	4	5	116	12	0	299
73								468	33	75	3	0	579
76									690	301	9	0	1000
79										1892	85	0	1977
90											1	0	1
99												144	144

TOTAL VALID COMPARISONS	12600

Coding Guide

Code	Response
Correct Response	
10	Sandy's. The reflector reflects all the light towards the wall.
11	Sandy's. In Jim's flashlight the light shines in all directions.
12	Any combination of codes 10, 11.
19	Sandy's. Other correct explanations.
Incorrect Response	
70	Jim's.
71	Jim's but with an explanation that belongs to Sandy's.
72	Jim's with another explanation.
73	Jim's or Sandy's, with no explanation.
76	Merely repeats information in stem.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

New Species in Area

Item S17

What could be the unwanted consequences of introducing a new species to a certain area? Give an example.

	FREQUENCIES OF MATCHED CODES										
	20	**21**	**29**	**10**	**11**	**12**	**19**	**70**	**79**	**90**	**99**	**ROW SUM**
20	112	271	44	131	65	29	52	17	20	2	0	743
21		275	53	242	170	22	106	63	53	3	0	987
29			26	111	39	65	63	24	39	2	0	369
10				1495	57	467	508	80	380	37	2	3026
11					65	43	55	47	69	6	0	285
12						389	142	106	115	9	0	761
19							105	44	280	18	0	447
70								156	90	13	0	259
79									517	181	9	707
90										70	94	164
99											1381	1381

TOTAL VALID COMPARISONS 9129

Coding Guide	
Code	**Response**
Correct Response	
20	States that the natural (ecological) balance will be upset. A realistic example of a species is given.
21	States that the new species may take over and gives examples.
29	Other correct responses with examples.
Partial Response	
10	Adequate explanation (as in codes 20, 21), but no concrete and realistic example is given.
11	Only the realistic example is given, but no explanation.
12	States the new species cannot live here.
19	Other partially correct.
Incorrect Response	
70	Only an unrealistic example is given.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

BEST COPY AVAILABLE

APPENDIX I

DATA CLEANING AND CONSISTENCY CHECKS

CHECKS UNDERTAKEN AT THE DPC

This appendix contains a complete compilation of all checks undertaken at the DPC. Each check is presented in the following way:

Problem Number (Affected files): Description of the problem

→ Applied cleaning rule: Corrections have been undertaken

⊗ Applied cleaning rule: A warning has been given, but no corrections have been undertaken.

Please Note: The cleaning rule reports the action undertaken at the DPC if no solution could be found by inspecting the instruments and forms and if no instructions on how to handle these cases were given by the national center.

COLUMN SHIFT CHECK

This cleaning step identifies potential column shifts. It identifies all observations in which valid data have been found in the CHECKn variables, which should be blank.

101 (CG, TG, SA, SG, SR): The listed observations include one CHECKn variable that contains a value.
→ If the reason for the column shift could not be identified, all data between the effected CHECKn variable and the preceding one have been coded to 'Not administered'.

102 (CG, TG, SA, SG, SR): The listed observations include more than one CHECKn variable that contains values.
→ If the reason for the column shifts could not be identified, all data between the last correct CHECKn variable and this one have been coded to 'Not administered'.

IDENTIFICATION CHECK - STUDENT FILE

This set of cleaning steps identifies all problems with IDs within an observation. This includes deviations from the hierarchical ID system described in the Main Study Manual as well as missing or incomplete identification.

201 (SA, SG, SR): The listed observations include a deviation from the hierarchical ID system. The Class ID, Student ID, and, if applicable, Teacher IDs all match, but all are inconsistent with the School ID.
→ The School ID has been made consistent with the other IDs.

202 (SA, SG, SR): The listed observations include a deviation from the hierarchical ID system. The School ID, Student ID, and, if applicable, all Teacher IDs match, but all are inconsistent with the Class ID.
→ The Class ID has been made consistent with the other IDs.

203 (SA, SG, SR): The listed observations include deviations from the hierarchical ID system. The Class ID does not match the class-specific part of the Student ID.
→ If it was not possible to identify the class to which the student belongs, the student has been treated as nonparticipating and has been assigned a weight of zero.

204 (SA, SG, SR): The listed observations include a deviation from the hierarchical ID system. The School ID, Class ID, and, if applicable, all Teacher IDs match, but all are inconsistent with the Student ID.
→ The Student ID has been made consistent with the other IDs.

205 (SG): The listed observations include a deviation from the hierarchical ID system. The School ID, Class ID, Student ID, and all but one of the Teacher IDs match.
→ For the nonmatching ID, the school-specific part of the Teacher ID has been made consistent with the other IDs.

206 (SG): The listed observations include Teacher IDs (ILTEACH1-6) that have a different number of digits from the other IDs.
→ If the correct Teacher ID could not be identified, the Teacher ID has been dropped and the corresponding variables ILTEACHx and ILLINKx have been coded to 'Not administered' (99998) and (98) respectively.

207 (SG): The listed observations include Teacher IDs (ILTEACH1-6) whose corresponding Teacher Link Numbers (ILLINK1-6) have been coded to 'Not administered' (98).
→ If no Teacher Link Number could be identified from the *Teacher Tracking Forms*, a new Teacher Link Number has been assigned. In this unlikely case, all course-related data would be lost and only person-related data would be available.

208 (SA, SG, SR):): The listed observations include more than one deviation from the hierarchical ID system.
→ If it was not possible to recreate the correct identification, the student has been treated as nonparticipating and a weight of zero has been assigned.

IDENTIFICATION CHECK - SCHOOL FILE

This set of cleaning steps identifies all inconsistencies between identification variables within the School File. This includes all inconsistencies between the IDs of replaced schools and originally sampled schools, as well as their Participation Indicators. Strata with only one school have also been flagged.

301 (CG): The listed schools are indicated as replacement schools (ILREPLAC contains a valid ID), but their Participation Indicators (ITPARTx) show that they are participating as originally sampled schools.
→ ITPARTx has been recoded to 'First replacement school' (1) or 'Second replacement school' (2).

302 (CG): The listed schools are indicated as replacement schools (ILREPLAC contains a valid ID).
→ The replaced school indicated by ILREPLAC has been entered.

303 (CG): The listed schools are indicated as having been replaced by other schools (their IDs were found in ILREPLAC for other schools), but their Participation Indicators (ITPARTx) are different from 'Nonparticipating' (0).
→ If data were entered for a corresponding school, the corresponding entry for ILREPLAC has been recoded to 'Not administered' (998). If no data were entered for a corresponding school, the Participation Indicator has been recoded to 'Nonparticipating' (0).

304 (CG): The listed schools are indicated as being 'Nonparticipating' (0), but no replacement schools have been entered.
⊗ A warning has been given, but no corrections have been undertaken.

305 (CG): The listed schools belong to strata which contain only one school.
⊗ A warning has been given, but no corrections have been undertaken.

IDENTIFICATION CHECK - TEACHER FILE

This set of cleaning steps identifies all problems with IDs within an observation. This includes deviations from the hierarchical ID system described in the Main Study Manual as well as missing or incomplete identification. Please note that all recoding has been performed under consideration of its consistency with other files.

401 (TG): The listed observations include a deviation from the hierarchical ID system. The Teacher ID and Class IDs all match, but all are inconsistent with the School ID.
→ The School ID has been made consistent with the other IDs.

402 (TG): The listed observations include a deviation from the hierarchical ID system. The School ID and Class IDs all match, but all are inconsistent with the Teacher ID.
→ The Teacher ID has been made consistent with the other IDs.

403 (TG): The listed observations include a deviation from the hierarchical ID system. The School ID, Teacher ID, and Class IDs all match, but all are inconsistent with one of the Class IDs.
→ This Class ID has been made consistent with the other IDs.

404 (TG): The listed observations include a Teacher ID whose length is inconsistent with that of the other IDs.
→ A new consistent Teacher ID has been assigned.

405 (TG): The listed observations include a Class ID whose length is inconsistent with that of the other IDs.
→ This Class ID has been made consistent with the other IDs.

406 (TG): The listed observations include more than one deviation from the hierarchical ID system.
→ The effected cases have been dropped from the file.

407 (TG): The listed observations do not include Teacher Link Numbers.
→ A dummy link number has been assigned. Teacher-student linkage is not possible.

408 (TG): The listed observations include a Class ID twice.
→ The redundant entry has been set to 'Not applicable' (99998).

409 (TG): The listed observations do not include Class IDs.
⊗ A warning has been given, but no corrections have been undertaken.

410 (TG): The listed observations have a different subject (IDSUBJCT) from the booklet administered. Only applicable to Population 2.
→ The Subject ID has been made consistent with the administered questionnaire.

DUPLICATE ID CHECK

This set of cleaning steps identifies all problems with IDs between observations. This includes duplicate IDs or records and inconsistent IDs within homogenous groups; e.g., inconsistent grade identification within a class.

501 (CG, TG, SA, SR, SG): The listed observations contain exactly the same IDs and data.
→ One of the observations has been dropped.

502 (CG, TG, SA, SR, SG): The listed observations contain exactly the same IDs, but different data.
→ A dummy ID has been be created for at least one of the observations.

504 (CG, TG, SA, SR, SG): The listed observations include a Population ID (IDPOP) that does not match the file.
→The Population ID has been corrected to agree with the source file.

505 (CG, TG, SA, SR, SG): The listed observations belong to the same school, but their Stratum IDs are different.
→ The Stratum ID has been recoded to agree with the majority of entries.

506 (SG): The listed observations belong to the same class, but their Grade IDs are different.
→ The Grade ID has been recoded to agree with the majority of entries and the teacher data.

507 (TG): The listed teachers teach the same classes, but the Grade IDs are different.
→ The Grade ID has been recoded to agree with the majority of entries and the student data.

508 (SG): The listed observations belong to the same class, but their Language IDs are different.
→ The Language ID has been recoded to agree with the majority of entries and the student data.

509 (SG): The listed observations belong to the same class, but their Stream IDs are different.
→ The Stream ID has been recoded to agree with the majority of entries and the student data.

510 (SG): The listed observations belong to the same class, but their Test Administrator IDs (ITADMINI) are different.
⊗ A warning has been given, but no corrections have been undertaken.

TEST INDICATOR CHECK

This set of cleaning steps identifies inconsistencies between the Test Indicator Variables (TOKENxx) and the data variables. The code '3' should be assigned if the corresponding instrument was administered. All other codes indicate that the instrument was not administered or that the instrument has been lost. Please note that the TOKENxx variables represent the availability of data. The information given with the cleaned TOKENxx variables has been transcribed into the Participation Indicators ITPARTx (which summarizes the availability of data, the participation status, and the exclusion status) after all cleaning steps were finalized. Therefore, the TOKENxx variables became redundant after cleaning and have been removed from the files.

601 (CG, TG, SG, SA, SR): The listed observations include at least one TOKENxx variable that is coded invalidly and less than five valid data have been found.
→ The TOKENxx variable has been recoded to 'Nonparticipating' (2 for teachers, 1 for other respondents).

602 (CG, TG, SG, SA, SR): The listed observations include at least one TOKENxx variable that is coded invalidly, but more than five valid data have been found.
→ The TOKENxx variable has been recoded to 'Participating' (3).

603 (CG, TG, SG, SA, SR): The listed observations include at least one TOKENxx variable that is coded to 'Participating' (3) and less than 3 corresponding data variables are different from 'Not administered'.
→ The TOKENxx variable has been recoded to 'Nonparticipating' (2 for teacher, 1 for other respondents).

604 (CG, TG, SG, SA, SR): The listed observations include at least one TOKENxx variable that is coded to 'Not participating' (2) for teachers, (1) for all others and more than ten valid data have been found.
→ The TOKENxx variable has been recoded to 'Participating' (3).

605 (SG, SA, SR): The listed observations include at least one TOKENxx variable that has been coded to 'Booklet lost' (2), but valid data have been found.
→ The corresponding TOKENxx variable has been recoded to 'Participating' (3).

606 '(CG, TG, SG, SA, SR): The listed observations include at least one TOKENxx variable that is coded to 'Participating' (3) but only missing data (different from 'Not administered') have been found.
⊗ These cases should be checked carefully. If NRCs did not indicate changes to these cases, it is assumed that the respondent returned the questionnaire but left everything blank, e.g., because he/she did not cooperate. A warning has been given, but no corrections have been undertaken.

610 (SG, SA, SR): The listed students participated in only one testing session.
⊗ These cases have been flagged because this can be an indication of data loss and should be checked. No corrections are necessary.

611 (SG, SA, SR): The listed students did not participate in a testing session. These cases should not be entered into the Student Achievement File or the Reliability File.
⊗ These cases have been flagged. Because dummy records have been created for all nonparticipating students, a warning has been given, but no corrections have been undertaken.

612 (CG): The listed schools are indicated as 'Nonparticipating' (0) by the Participation Indicator Variable ITPARTx, but valid data have been found.
→ The Participation Indicator has been recoded to 'Participating' (3).

613 (CG): The listed schools are indicated as 'Participating' (1, 2, or 3) by the Participation Indicator Variable ITPARTx, but no valid data have been found and the corresponding TOKENxx variable indicates that the school did not participate.
→ The TOKENxx or ITPARTx variable has been recoded to agree with to the student and teacher data for this school.

BOOKLET ID CHECK

This set of cleaning steps identifies inconsistencies between the Booklet ID (IDBOOK) and the corresponding Test Indicator Variables (TOKENxA, TOKENxB). The Booklet ID indicates which booklet has been administered to the students. The corresponding TOKENxx variables should be coded to '2' or '3' and all other TOKENxx variables should be coded to '1'. Please note that the TOKENxx were cleaned before, which means they represent the existence of data. All mispunches of TOKENxx have been corrected during the

Test Indicator Check; data are available. If a TOKENxx is different from '3', there are no data available.

701 (SA, SR): The listed observations include a Booklet ID that has an invalid code, and no TOKENxx variables are coded to '3'.
→ IDBOOK has been coded to 'No booklet assigned' (0).

702 (SA, SR): The listed observations include a Booklet ID that has an invalid code, and one TOKENxx variable is coded to '3'.
→ IDBOOK has been made consistent with the TOKENxx variable, i.e. it has been coded to agree with the data found in the record.

703 (SA, SR): The listed observations include a Booklet ID that has an invalid code, and two TOKENxx variables corresponding to one booklet are coded to '3'.
→ IDBOOK has been made consistent with the TOKENxx variable, i.e. it has been coded to agree with the data found in the record.

704 (SA, SR): The listed observations include a Booklet ID that has an invalid code, and two or more TOKENxx variables corresponding to different booklets are coded to '3'.
→ All data have been recoded to 'Not administered' and IDBOOK has been coded to 'No booklet assigned' (0).

705 (SA, SR): The listed observations include a Booklet ID that has a valid code, but one (and only one) TOKENxx variable (which does not correspond to the appropriate booklet) are coded to '3'.
→ IDBOOK has been made consistent with the data.

706 (SA, SR): The listed observations include a Booklet ID that has a valid code, but two TOKENxx variables (which correspond to the same booklet) are coded to '3'. The booklet to which the two TOKENxx variables correspond is different from the one indicated by IDBOOK.
→ IDBOOK has been made consistent with the data.

707 (SA, SR): The listed observations include a Booklet ID that has a valid code, but TOKENxx variables, which correspond to a different booklet from that indicated by IDBOOK, have been identified. The data conflict as to which booklet was administered.
→ All data have been recoded to 'Not administered'; the data corresponding to IDBOOK have been marked as 'Lost'.

LINKAGE CHECK STUDENT GENERAL FILE -> STUDENT ACHIEVEMENT FILE

This set of cleaning steps identifies inconsistencies between the Exclusion Indicator IDEXCLUD, the Participation Indicators ITPART1 (first testing session), ITPART2 (second testing session), ITPART3 (Student General File), and the corresponding TOKENxx variables. Therefore, the linkage between the Student General File and the Student Achievement File will also be checked. Before correcting inconsistencies between TOKENxx variables in one file and Participation Indicator Variables ITPART1 and ITPART2 in the other, the ITPARTx linkage has to be checked carefully.

1001 (SA, SG): The listed observations include Participation Indicators ITPARTx that identify the particular student as 'Participating' (3), but no matching data have been found in the Student Achievement File. Either

the Participation Indicator is coded incorrectly or the linkage between the Student General File and the Student Achievement File is incorrect.
→ A dummy record has been created for the achievement data, which have been marked as 'Lost'; the corresponding TOKENxx variables have been coded to 'Booklet/data lost' (2).

1002 (SA, SG): The listed observations include the Participation Indicators ITPARTx that identify the particular student as 'Nonparticipating' (0, 1, or 2), but a matching record, which contains valid data, has been found in the Student Achievement File. Either the Participation Indicator ITPARTx is coded incorrectly or the linkage between the Student General File and the Student Achievement File is incorrect.
→ The Participation Indicator ITPARTx has been made consistent with the data, i.e., it has been coded to 'Participating' (3).

1003 (SA, SG): The listed observations include a record in the Student Achievement File for which all TOKENxx for the tested booklet are coded to 'Nonparticipating' (1) and for which the corresponding ITPARTx variables identify the student as 'Nonparticipating' (0, 1, or 2). Records without data should not be entered into the Student Achievement File.
⊗ No corrections are necessary.

1004 (SA, SG): The Participation Indicators ITPARTx for the listed observations identify the student as 'Nonparticipating' (0, 1, or 2) and no valid data have been found. At least two of the Participation Indicators ITPARTx mark the student as excluded, but the Exclusion Indicator IDEXCLUD is coded to 'Not excluded' (9).
→ The Participation Indicators ITPARTx have been recoded to 'Student absent' (2) instead of 'Excluded' (0).

1005 (SA, SG): The Participation Indicators ITPARTx for the listed observations identify the student as 'Participating' (3) and valid data have been found, but the Exclusion Indicator IDEXCLUD is coded differently from 'Not excluded' (9).
→ The Participation Indicators ITPARTx have been recoded to 'Participating, but excluded' (5).

1006 (SA, SG): The Participation Indicators ITPARTx for the listed observations identify the student as 'Nonparticipating' (0, 1, or 2) and no valid data have been found, but the Exclusion Indicator IDEXCLUD has an invalid code.
⊗ A warning has been given, but no corrections have been undertaken.

1007 (SA, SG): All but one of the Participation Indicators ITPARTx for the listed observations are consistent both with the data found and with the Exclusion Indicator IDEXCLUD. Valid data have been found for the corresponding session.
⊗ The Participation Indicator ITPARTx has been made consistent with the data.

1008 (SA, SG): All but one of the Participation Indicators ITPARTx for the listed observations are consistent both with the data found and with the Exclusion Indicator IDEXCLUD. One Participation Indicator ITPARTx is coded to 'Nonparticipating' and the corresponding booklet is indicated by TOKENxx as being 'Lost' (2).
→ The TOKENxx variable has been coded to 'Nonparticipating' (1).

1009 (SA, SG): All but one of the Participation Indicators ITPARTx for the listed observations are consistent both with the data found and with the Exclusion Indicator IDEXCLUD. No data have been found for the corresponding session, but the booklet is not indicated by TOKENxx to be 'Lost' (2).
→ The Participation Indicator ITPARTx has been coded to 'Booklet/data lost' (4).

1010 (SA, SG): The listed observations include inconsistencies between Participation Indicators. Different codes were given for nonparticipation (0 = 'Student excluded', 1 = 'Student left school', or 2 = 'Student absent').
→ These codes have been made consistent. Either 'Student left school' or 'Student absent' was used twice and the third Participation Indicator has been made consistent, or all Participation Indicator Variables which indicate nonparticipation have been recoded to 'Student absent' (2). If a student was excluded and no valid data have been found, all Participation Indicators have been recoded to 'Student excluded' (0).

1011 (SA, SG): The listed observations include several inconsistencies between Participation Indicators, Exclusion Indicators, and data.
→ Inconsistent indicators have been made consistent with the data. Participation indicators ITPARTx and TOKENxx have been recoded to 'Participating' (3) if data were found. ITPARTx has been coded to 'Student absent in session' if no data were found.

1012 (SA, SG): The listed observations include inconsistencies between the Class ID (School ID) in the Student Achievement File and the Class ID (School ID) in the Student General File.
→ As with many other problems, this problem creates at least two messages in the cleaning report. The rules for the Identification Check, where these cases have also been reported, have been applied.

LINKAGE CHECK RELIABILITY FILE -> STUDENT ACHIEVEMENT FILE

This set of cleaning steps identifies all inconsistencies between the Reliability Coding File and the corresponding Student Achievement File.

1111 (SA, SR): The listed observations in the Student Achievement File do not match those in the corresponding Reliability Coding File.
→ The affected cases have been dropped from the Reliability Coding File.

1112 (SA, SR): The Reliability Coding Files and the Student Achievement Files for the listed observations indicate that different booklets were administered.
→ The affected cases have been dropped from the Reliability Coding File.

1113 (SA, SR): The listed observations are missing some data for the Reliability Coding File, although the Student Achievement File contains corresponding data.
⊗ No corrections could be performed.

1114 (SA, SR): The booklet set (indicated by ITBSET in the Reliability Coding File) for the listed observations' is different from the booklet set indicated in the Student Achievement File.
⊗ No corrections could be performed. It should be checked whether the data were obtained from the same booklets.

1115 (SA, SR): The listed observations indicate the same coder for both the Reliability Coding File and the Student Achievement File.
⊗ No corrections could be performed. It should be checked if the reliability coding followed the correct procedures.

1116 (SA, SR): The listed observations include inconsistencies between the Class ID (School ID) within the Student Achievement File and the Class ID (School ID) in the Reliability Coding File.
→ This problem creates at least two messages in the cleaning report. The rules for the Identification Check, where these cases have also been reported, have been applied.

1121 (SA, SR): The listed observations include a record in the Student Achievement File, which corresponds to one in the Reliability Coding File, which contains invalid data (all booklets have been lost or the student did not participate). These students should not be selected for reliability coding.
→ The affected cases have been dropped from the Reliability Coding File.

LINKAGE CHECK STUDENT ACHIEVEMENT FILE -> STUDENT GENERAL FILE

This cleaning step identifies records in the Student Achievement File which can not be merged to the Student General File.

1201 (SA, SG): The listed observations do not include any matching observations within the Student General File.
→ A dummy record has been created for the background data and all Participation Indicators have been made consistent with the data, i.e., the indicators corresponding to the background session have been coded to 'Booklet lost'. Indicators that could be derived from the achievement data, e.g., the Grade ID from the Class ID, have been corrected.

STUDENT - TEACHER LINKAGE CHECK

This set of cleaning steps identifies linkage problems between the Teacher and Student Files.

1311 (SG, TG): The listed observations include a combination of Teacher ID and Teacher Link Number which can not be found in the Teacher Files.
→ A dummy teacher has been created so that complete linkage is possible. All indicators in the dummy teacher record have been set to 'Teacher did not participate'.

1312 (SG, TG): The listed observations include a class linked to the teacher (in ILCLASS1-3) which can not be found in the Student Files.
→ The affected variable ILCLASSx has been recoded to 'Not administered'.

1313 (SG, TG): The listed observations include a Grade ID that is inconsistent between the teacher and the students linked to him/her.
→ The Grade IDs have been made consistent for students and teachers to agree with the majority of data.

1314 (SG, TG): The following teachers' students are from classes which were not given in ILCLASS1, ILCLASS2, or ILCLASS3.
→ All class IDs corresponding to classes in which students are linked to the particular teacher will be added to ILCLASS1-3.

1315 (SG, TG): The following teachers are linked to classes (ILCLASS1-3), which are represented in the datafiles, but no student is linked to the corresponding teacher. (Variables ILTEACH1-6, ILLINK1-6)
→ IDs have been made consistent.

1321 (SG): No math teacher is indicated for the listed students.
⊗ No corrections could be performed.

1322 (SG): No science teacher is indicated for the listed students.
⊗ No corrections could be performed.

1323 (SG): No teacher is linked to the listed students.
⊗ No corrections could be performed.

SCHOOL - STUDENT LINKAGE CHECK

This set of cleaning steps identifies linkage problems between the Student and School Files.

1411 (CG, SG): The listed schools are marked as 'Not participating' (0), but student data have been found in the Student File.
→ The participation status has been coded to 'Participating' (3) to agree with the data found.

1412 (CG, SG): The listed schools are indicated to be 'Participating' (1, 2, or 3), but no student data have been found in the Student File.
⊗ No corrections could be performed.

1413 (CG, SG): The listed schools have a different Stratum ID than the students linked to them.
⊗ The Stratum IDs have been made consistent for the school, the teacher, and all students, to agree with the majority of data.

1414 (CG, SG): The data for the listed schools includes more students than are enrolled in the schools (ITMOS).
⊗ No corrections could be performed.

1421 (SG): The listed observations are linked to schools which have not been found in the School File.
→ A dummy school has been created.

SCHOOL - TEACHER LINKAGE CHECK

This set of cleaning steps identifies linkage problems between the Teacher and School Files.

1511 (CG, TG): The listed schools have different Stratum IDs from the teachers linked to them.
→ The Stratum IDs have been made consistent for the school, the teacher, and all students to agree with the majority of data.

1512 (TG): The listed schools are marked as 'Not participating' (0), but teacher data have been found in the Teacher File.
→ Corrections have been undertaken, depending on whether students were found for the corresponding schools or not.

1513 (TG): The listed schools are indicated to be 'Participating' (1, 2, or 3), but no teacher data have been found in the Teacher File.
⊗ Corrections have been undertaken, depending on whether students were found for the corresponding schools or not.

1521 (TG): The listed observations are linked to schools which have not been found in the School File.
⊗ Corrections have been undertaken, depending on whether students were found for the corresponding schools or not.

RANGE VALIDATION CHECK

This cleaning step identifies variables with values which do not match the range validation criteria as specified in the electronic codebook (structure database).

2001 The listed observations include noncategorical variables with values which do not match the range validation criteria.
→ These variables have been set to 'Invalid'.

2002 The listed observations include categorical variables with values which do not match the range validation criteria.
→ These variables have been set to 'Invalid'.

STUDENT SPLIT VARIABLE CHECK

In some question blocks, students have been allowed to answer using one of two response options per question (e.g., 'Yes' or 'No' to questions on home possessions). These question blocks have been coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem. Respondents often do not use the 'No' option and mark only options with 'Yes'. For questions with a large number of options, it is assumed that in case, in which no 'No' option was used but 'missing' was coded, the respondent meant 'No'. This algorithm can not work for questions with only a few items.

2101 (SG): The listed cleaning step identifies all problems in the question block 'Do each of these people live at home with you most or all of the time?
→ If at least one of the dichotomous variables has been coded to 'Yes' (1) and all other variables have been coded to 'Missing', then all variables coded to 'Missing' have been recoded to 'No'.

2102 (SG): The listed cleaning step identifies all problems in the question block 'Do you have any of these items at home?
→ If at least one of the dichotomous variables has been coded to 'Yes' (1) and all other variables have been coded to 'Missing', then all variables coded to 'Missing' have been recoded to 'No' (2).

TEACHER SPLIT VARIABLE CHECK

The following problems have been combined in this category:

In some question blocks, teachers were allowed to answer using one of two response options per question (e.g., 'Yes' or 'No' to questions on grades taught). These question blocks were coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem. Respondents often do not use the 'No' option and mark only options with 'Yes'. For questions with a large number of options, it is assumed that in case, in which no 'No' option was used but several times 'missing' was coded, the respondent meant 'No'. This algorithm can not work for questions with only a few items.

In some question blocks, lists were given (e.g., in the Section 'Opportunity to Learn'). The elements of the lists could be' Checked' or 'Not checked'. Distinguishing between 'Not checked' and 'Missing/nonresponse' is not possible. It is assumed, that in cases where 'Missing' was coded and at least one option was checked, 'Not checked' should have been coded.

In other question blocks, teachers were asked to assign numbers (e.g., order or time). Distinguishing between '0' and 'Missing/nonresponse' is a problem. Respondents often do not use the '0' option and enter only numbers for options which are applicable. For questions with a large number of options, it is assumed that if no '0' option was used but several times 'missing' was coded, the respondent meant '0'. This algorithm can not work for questions with only a few items.

2201 (TG): In the question blocks concerning the grades taught, teachers were allowed to answer using one of two response options per question ('Yes' or 'No'). These question blocks were coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem.
→ If at least one of the dichotomous variables are coded to 'Yes' (1) and all other variables are coded to 'Missing' (9), then all variables coded to 'Missing' have been recoded to 'No'(2).

2202 (TG): In the question block concerning the textbooks used, teachers were allowed to answer using one of two response options per question ('Yes' or 'No'). These question blocks were coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem.
→ If at least one of the dichotomous variables are coded to 'Yes' (1) and all other variables are coded to 'Missing' (9), then all variables coded to 'Missing' have been recoded to 'No'(2).

2203 (TG): In the question block concerning the topics in the last lesson (A/BTBMTOxx and BTBSTOxx), teachers were allowed to answer using one of two response options per question ('Yes' or 'No'). These question blocks were coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem.
→ If at least one of the dichotomous variables is coded to 'Yes' (1) and all other variables are coded to 'Missing' (9), then all variables coded to 'Missing' have been recoded to 'No'(2).

2204 (TG): Teachers were asked to list the activities of the last class hour in order.
→ If at least one number is assigned, but no 0 has been detected, all 'Missing' have been recoded to 0.

2206 (TG): Teachers were asked to assign times to the activities in the last lesson. They should enter 0 if they did not do a certain activity
→ If at least one number is assigned, but no 0 has been detected, all 'Missing' have been recoded to 0.

2208 (TG): This cleaning step identifies inconsistencies in the lists concerning the opportunity to learn a topic. (Population 2: Section C) For each topic, two lists were given, one which should have been answered if the topic was taught and the second which should have been answered if the topic was not taught. Both lists have been checked separately for internal consistency
→ If at least one variable in a list is coded to 'Checked' (2), all variables coded to 'Missing' (9) have been recoded to 'Not checked' (1)
→ If no variables in a list are coded to 'Checked' (2), all variables coded to 'Not checked' (1) have been recoded to 'Missing' (9).

SCHOOL SPLIT VARIABLE CHECK

The listed problems have been combined in this category: In some question blocks, principals were allowed to answer using one of two response options per question (e.g., 'Yes' or 'No' to questions on grades found in the school). These question blocks were coded using dichotomous variables. Distinguishing between 'Missing/nonresponse' and 'No' is a problem. Respondents often do not use the 'No' option and mark only options with 'Yes'. For questions with a large number of options, it is assumed that if no 'No' option was used but 'missing' was coded, the respondent meant 'No'. This algorithm can not work for questions with only a few items.

In other question blocks, principals were asked to assign numbers (e.g., time). Distinguishing between '0' and 'Missing/nonresponse' is a problem. Respondents often do not use the '0' option and enter only numbers for options which are applicable. For questions with a large number of options, it is assumed that if no '0' option was used but several times 'missing' was coded, the respondent meant '0'. This algorithm can not work for questions with only a few items.

2301 (CG): This cleaning step identifies all problems in the question block 'Which of following grade levels are found in your school?'
→ If at least one of the dichotomous variables is coded to 'Yes' (1) and all other variables are coded to 'Missing' (9), then all other variables coded as 'Missing' (9) have been recoded to 'No' (2).

2302 (CG): This cleaning step identifies all cases in which the respondent was asked to write a 0 for 'None', but he/she ignored this and left all fields blank. The listed variables have been checked:(1) number of full-time equivalents A/BCGFTE1-6;(2) principal's activities A/BCGAC01-14;(3) percentage of students coming from different background (A/BCBGSTD1-8.)
→ In all cases in which no '0', one or more 'Missing', and valid answers were given, all 'Missing' have been recoded to '0'.

2303 (CG): The listed cleaning step identifies all problems in the question block 'On what basis are pupils admitted to your school ?' (Population 1 SCQ1-29x, Population 2 SCQ2-30x)
→ If at least one of the dichotomous variables is coded to 'Yes' (1) and all other variables are coded to 'Missing' (9), then all other variables coded as 'Missing' (9) have been recoded to 'No' (2).

STUDENT FILTER VARIABLE CHECK

These cleaning steps identify all problems between filter variables and dependent variables in the Student File(s). In a first cleaning step, the filter variable has been made consistent with the dependent variable(s). In a second cleaning step all dependent variables coded to 'Missing' have been recoded to 'Logically not applicable' if the filter variable was coded to 'No'.

2401 (SG): This cleaning step identifies all problems between the statement concerning 'Born in country' and 'Age when student came to country. ('Note: this check requires the following checks (I) Check of consistency in 'Age of Student' obtained from tracking information and questionnaire information. (II) Check of consistency in 'Date of Testing'. (III) Check of consistency between 'Age of student' and 'Age when student came to country'.)
→ If the student's age is valid but the student indicated in the filter question that he/she was born in country (1), then the filter variable has been recoded to 'No' (2).

2402 (SG): The listed cleaning step identifies all problems between the statement concerning 'Born in country' and 'Age when student came to country. (Note: this check requires the following checks: (I) Check of consistency in 'Age of Student' obtained from tracking information and questionnaire information. (II) Check of

consistency in 'Date of Testing'.(III) Check of consistency between 'Age of student' and 'Age when student came to country'.
→ If the student's age when he/she came to the country was invalid and the student indicated in the filter question that he/she was 'Born in country' (1), then the student's age when he/she came to the country has been recoded to 'Logically not applicable' (96).

2403 (SG): This cleaning step reports all inconsistencies between the filter question concerning <SCIENCE SUBJECT> and completion of questions in the corresponding part of the *Student Questionnaire*. (Note: <BIO> refers to the biology part, <CHE> refers to the chemistry part, <EAR> refers to the earth science part, <PHY> refers to the physics part.
→ If the student states in the filter question that he/she does not study <SUBJECT> this year (1), but answers more than 3 questions concerning the <LESSON>, the filter has been recoded to 'Yes' (2).
→ If the student states in the filter question that he/she does not study <SUBJECT> this year (1), but answers more than 2 questions concerning the <SUBJECT>, a warning has been given, but no corrections have been undertaken.

2404 (SG): The listed cleaning step identifies all inconsistencies between the filter question concerning <SUBJECT> and completion of less than 3 questions in the corresponding part of the *Student Questionnaire*. Note: <BIO> refers to the biology part, <CHE> refers to the chemistry part, <EAR> refers to the earth science part, <PHY> refers to the physics part.
→ If variables coded differently from 'Not administered' (8) have been found in the part under consideration, the variables have been recoded to 'Logically not applicable' (B).

TEACHER FILTER VARIABLE CHECK

This cleaning steps identifies all problems between filter variables and dependent variables in the Teacher Files. In a first cleaning step the filter variable has been made consistent with the dependent variable(s). In a second cleaning step all dependent variables coded to 'Missing' have been recoded to 'Logically not applicable' if the filter variable was coded to 'No'.

2501 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether <SUBJECT> is taught this year and dependent variables which indicate at which grade levels <SUBJECT> is taught. (Note: <SUBJECT> can be mathematics or science.)
→ If the filter indicates that NO <SUBJECT> is taught (2), but at least one grade level variable indicates that <SUBJECT> is taught, then the filter has been recoded to 'Taught' (1).
→ If the filter indicates that <SUBJECT> is taught (1), but NO grade level variable indicates that <SUBJECT> is taught, then the filter has been recoded to 'Not taught' (2).

2502 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether <SUBJECT> was taught this year and the number of hours formally scheduled for teaching it. (Note: <SUBJECT> can be mathematics or science.)
⊗If the filter indicates that NO <SUBJECT> is taught (2), but time is officially scheduled to teaching it, a warning has been given, but no corrections have been undertaken.

2521 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether <SUBJECT> was taught this year and the type of questionnaire which was administered for <SUBJECT>. (Note: <SUBJECT> can be mathematics or science.)

⊗ If the filter indicates that NO <SUBJECT> is taught (2), but a corresponding questionnaire is completed, a warning has been given, but no corrections have been undertaken.

2522 (TG): This cleaning step identifies inconsistencies between two filters which indicate that neither mathematics nor science is taught this year.

⊗ If both filters indicate that NO MAT and NO SCI is taught, a warning has been given, but no corrections have been undertaken.

2503 (TG): This cleaning step identifies inconsistencies between a filter which indicates that no <SUBJECT> is taught this year and dependent variables which indicate at which grade levels it is taught. Also, the identification variable IDSUBJCT has been included.(Note: <SUBJECT> can be mathematics or science.)

⊗ If the filter indicates that NO <SUBJECT> is taught (2), and NO grade level variables indicate that it is taught, but IDSUBJCT indicates that it is taught in the selected class, a warning has been given, but no corrections have been undertaken.

2531 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether <SUBJECT> is taught this year and dependent variables which indicate at which grade levels <SUBJECT> is taught. (Note: <SUBJECT> can be mathematics or science.)

→ If the filter indicates that NO <SUBJECT> is taught (2), and the dependent variables were coded to 'Missing' (9) or 'Not taught' (2), the dependent variable has been recoded to 'Logically not applicable' (B).

2504 (TG): This cleaning step identifies inconsistencies between a variable which indicates whether a textbook is used in <SUBJECT> and dependent variables which indicate which textbooks are used. (Note: <SUBJECT> can be mathematics or science.)

→ If it is indicated that NO textbook is used (2), but textbooks are checked in the list or a name is entered, then the filter has been recoded to 'Yes' (1).

2505 (TG): This cleaning step identifies inconsistencies between a filter variable which indicates whether a textbook is used in<SUBJECT> and dependent variables which indicate which textbooks are used.(Note: <SUBJECT> can be mathematics or science.)

→ If the filter indicates that NO textbook is used (2) and dependent variables were coded to 'Missing', then dependent variables have been recoded to 'Logically not applicable'.

2506 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether a textbook is used in <SUBJECT>and a dependent variable which indicates whether teaching is based on a textbook. (Note: <SUBJECT> can be mathematics or science.)

⊗ If it is indicated that NO textbook is used, but teaching is based on textbooks, then a warning has been given, but no corrections have been undertaken.

2507 (TG): This cleaning step identifies inconsistencies between a filter variable which indicates whether a textbook is used in<SUBJECT> and a dependent variable which indicates whether teaching is based on a textbook. (Note: <SUBJECT> can be mathematics or science.)

→ If it is indicated that NO textbook is used and teaching is not based on textbooks (Missing), then 'Missing' has been recoded to 'Logically not applicable' (B).

2508 (TG): This cleaning step identifies inconsistencies between a filter which indicates whether homework was given in the recent <CLASS> and a dependent variable which indicates the time necessary to do the homework. (Note: <CLASS> can be the mathematics class or science class.)

→ If the filter indicates that NO homework was assigned (2), but a time was entered, the filter has been recoded to 'Yes' (1).

2509 (TG): This cleaning step identifies inconsistencies between a filter variable which indicates whether homework was given in the recent <CLASS> and a dependent variable which indicates the time necessary to do the homework. (Note: <CLASS> can be the mathematics class or the science class.)

→ If the filter indicates that NO homework was assigned (2) and the dependent variable is coded to 'Missing' (999), the dependent variable has been recoded to 'Logically not applicable' (996).

2510 (TG): This cleaning step identifies inconsistencies in a filter indicating whether or not a topic was taught and two dependent sets of variables ('Yes list' and 'No list').

⊗ If both lists are checked and the filter is set to 'No' (2), a warning has been given, but no correction has been undertaken.

→ If the 'Yes list' is checked, but not the 'No list', and the filter is set to 'No' (2), the filter has been recoded to 'Yes' (1).

→ If the 'No list' is checked, but not the 'Yes list', and the filter is set to 'Yes' (1), the filter has been recoded to 'No' (2).

2511 (TG): This cleaning step identifies inconsistencies in a filter indicating whether or not a topic was taught and two dependent sets of variables ('Yes list' and 'No list').

⊗ If both lists are checked and the filter is set to 'Yes' (1), a warning has been given, but no correction has been undertaken.

2512 (TG): This cleaning step identifies inconsistencies in a filter indicating whether or not a topic was taught and two dependent sets of variables ('Yes list' and 'No list').

→ If only one list is checked and the filter is consistent with the list checked, but variables in the other list are coded to 'Missing' (9), 'Missing' has been recoded to 'Logically not applicable' (B).

2513 (TG): This cleaning step identifies inconsistencies in a filter indicating whether or not science is taught mainly as a separate subject and two dependent variables indicating the time science is taught (one variable if the filter states that science is taught as a separate subject and one variable if the filter states that science is not taught as a separate subject).

⊗ If both times have been assigned, a warning is given, but no correction has been undertaken.

→ If the 'Yes'-time is assigned, and the filter is set to 'No' (2), the filter has been recoded to 'Yes' (1)

→ If the 'No'-time is assigned, and the filter is set to 'Yes' (1), the filter has been recoded to 'No' (2).

2514 (TG): This cleaning step identifies inconsistencies in a filter indicating whether or not science is taught mainly as a separate subject and two dependent variables indicating the time science is taught (one variable if the filter states that science is taught as a separate subject and one variable if the filter states that science is not taught as a separate subject)

→ If only one time is assigned and the filter is consistent with the assigned time, but the second time is coded to 'Missing' (999) or 0, the 'Missing' or 0 has been recoded to 'Logically not applicable' (996).

SCHOOL FILTER VARIABLE CHECK

This cleaning steps identifies all problems between filter variables and dependent variables in the School File. In a first cleaning step, the filter variable has been made consistent with the dependent variable(s). In a second cleaning step, all dependent variables coded to 'Missing' have been recoded to 'Logically not applicable' if the filter variable was coded negatively.

2601 (CG) This cleaning step identifies all cases in which the instructional time is the same for both the upper grade and the lower grade.
→ If the instructional time is the same for both grades, but different times were entered for the lower and the upper grade, the filter (A/BCBGINST) has been recoded to 'Not checked' (1).
→ If the instructional time is not the same for both grades according to the filter variable, but all times were entered identically for both grades, the filter (A/BCBGINST) has been recoded to 'Checked' (2).

2602 (CG): This cleaning step identifies all cases in which the school week is not divided into instructional periods according to the filter variable, but the dependent questions have been answered.
→ The filter variable A/BCBGDIVI has been recoded to 'Yes' (1).

2603 (CG): This cleaning step identifies all cases in which the school week is not divided into instructional periods according to the filter variable, but the dependent variables have been coded to 'Missing' instead of 'Not applicable'.
→ The dependent questions have been recoded to 'Logically not applicable'.

2604 (CG): This cleaning step identifies problems with filter and dependent questions for the remedial teaching of math or science and the special enriched teaching of math or science. The following filter questions and dependent variables have been checked: (1) variables on teaching remedial math: A/BCBMRMDL <-> A/BCBMRMD1-4; (2) variables on teaching remedial science: A/BCBSRMDL <-> A/BCBSRMD1-4; (3) variables on special enrichment activities in math: A/BCBMENRH <-> A/BCBMENR1-4;(4) variables on special enrichment activities in science: A/BCBSENRH <-> A/BCBSENR1-4.
→ If at least one dependent question was answered with 'Yes' (1), but the filter question was answered with 'No' (2), the filter question has been recoded to 'Yes' (1.)
→ If at least one dependent question was answered with 'No' (2) and none was answered with 'Yes' (1) and the filter question has been answered with 'No' (2), the dependent questions have been recoded to 'Logically not applicable' (B).

2605 (CG): This cleaning step identifies problems with filter and dependent questions for the upper grade courses in math or science. The listed filter and dependent variables have been checked: (1) variables on upper grade courses in mathematics: A/BCBMUSCO <-> A/BCBMUC1-62,A/BCBMUFC1-8; (2) variables on upper grade courses in science: A/BCBSUSCO <-> A/BCBSUC1-62,A/BCBSUFC1-8.
→ If the filter variable (A/BCMUSCO A/BCSUSCO) was coded to 'All students take the same courses' (1), but questions answered indicate that students take different courses, the filter variable has been recoded to 'Different courses' (2).
→ If the filter variable was coded to 'Different courses' (2), but the dependent variables indicate that students take the same courses, the filter variable has been recoded to 'All students take the same courses' (1).
⊗ Answers given for both blocks of dependent questions have been flagged.

2606 (CG): This cleaning step identifies all cases in which dependent questions were incorrectly coded as 'Missing' (9), but the filter question indicates that the questions should not be answered. The listed filter and dependent variables have been checked: (1) variables on teaching remedial math: A/BCBMRMDL <-> A/BCBMRMD1-4; (2) variables on teaching remedial science: A/BCBSRMDL <-> A/BCBSRMD1-4; (3) variables on special enrichment activities in math: A/BCBMENRH <-> A/BCBMENR1-4; (4) variables on special enrichment activities in science: A/BCBSENRH <-> A/BCBSENR1-4; (5) variables on upper grade courses in mathematics: A/BCBMUSCO <-> A/BCBMUC1-62,A/BCBMUFC1-8; (6) variables on upper grade courses in science: A/BCBSUSCO <-> A/BCBSUC1-62,A/BCBSUFC1-8.

→ Incorrectly coded dependent variables have been recoded to 'Logically not applicable' (B).

INCONSISTENCY CHECK-STUDENT FILE

This set of cleaning steps identifies all problems between data variables.

3101 (SG): This cleaning step identifies all problems with date of testing. Different or missing testing dates within a class have been detected.

→ If the year of testing is missing for the whole class, the year has been recoded to 1994 for countries in the southern hemisphere and to 1995 for countries in the northern hemisphere.

⊗ If the month of testing is missing for the whole class, a warning has been given, but no correction has been undertaken.

→ If the testing dates are missing for single students, they have been replaced by the values found for the other students in the class (if they are otherwise consistent).

⊗ If the date of testing differs for students within a class, a warning has been given, but no corrections have been undertaken.

3102 (SG): This cleaning step identifies problems with 'Date of birth' obtained from tracking information and questionnaire information.

⊗ If both tracking and questionnaire information are missing, a warning has been given, but no corrections have been undertaken.

3103 (SG): This cleaning step identifies problems with 'Date of birth' obtained from tracking information and questionnaire information. 'Dates of birth' which are 'Incomplete' but not 'Missing' in the questionnaire information and are 'Incomplete' or 'Missing' in the tracking information have also been reported.

→ If tracking information is available and questionnaire information is missing, then the questionnaire variables have been recoded to the tracking information.

⊗ If tracking information is available, but is different from questionnaire information (not missing), a warning has been given, but no corrections have been undertaken.

3104 (SG): This cleaning step identifies all problems with student's sex obtained from tracking information and from the *Student Questionnaires*.

→ If tracking information is available and questionnaire information is missing, the questionnaire information has been replaced by the tracking information and vice versa.

⊗ If the information is different in both sources (and not missing), a warning has been given, but no corrections have been undertaken.

3105 (SG) This cleaning step identifies all problems between 'Age of student' and 'Age when student came to country'. (Note: this check requires the following checks: (I) Check of consistency in 'Age of student' obtained from tracking information and questionnaire information;(II) Check of consistency in 'Date of testing'. If 'Age of student' can not be computed from the data because 'Date of test' is not available, '94 is used in southern hemisphere countries and '95 in northern hemisphere countries.
→ If the student is younger than indicated in 'Age when student came to country', then 'Age when student came to country' has been coded to 'Invalid' (97).

3106 (SG): This cleaning step identifies inconsistencies between 'Number of people living at home' and the question block 'Do each of these people live at home with you most or all of the time?' Students where the number of questions answered with 'Yes' exceeds the 'Number of people living at home' by one have been identified and counted. It is assumed, that they forgot to include themselves in the number of people living at home.
⊗ A warning has been given, but no corrections have been undertaken.

3107 (SG): This cleaning step identifies inconsistencies between 'Number of people living at home' and the question block 'Do each of these people live at home with you most or all of the time?' Students where the number of questions answered with 'Yes' exceeds the 'Number of people living at home' by more than one have been identified and counted.
⊗ A warning has been given, but no corrections have been undertaken.

3108 (SG): This cleaning step identifies all problems with <SUBJECT>: 'The teacher gives us homework' and the dependent questions concerning homework in <SUBJECT> lessons.(Note:<MAT> refers to mathematics lessons <SCI> refers to science lessons)
⊗ If at least one variable (but not all) concerning 'Homework in <SUBJECT> lessons' is coded differently from 'Never' (1 or 2), but the variable concerning 'Homework given in <SUBJECT> lessons' is coded to 'Never' (3), then a warning has been given, but no corrections have been undertaken.

3109 (SG) This cleaning step identifies all problems with <SUBJECT> 'The teacher gives us homework' and the dependent questions concerning homework in <SUBJECT> lessons. (Note:<MAT> refers to mathematics lessons <SCI> refers to science lessons)
→ If all variables concerning 'Homework in <SUBJECT> lessons' are coded differently from 'Never' (1 or 2), but the variable concerning 'Homework given in <SUBJECT> lessons' is coded to 'Never' (3), then the variable concerning 'Homework given in <SUBJECT> lessons' has been coded to 'Invalid' (I).

3110 (SG) This cleaning step identifies all problems with <SUBJECT>: 'The teacher gives us homework' and the dependent questions concerning homework in <SUBJECT> lessons. (Note:<MAT> refers to mathematics lessons <SCI> refers to science lessons)
⊗ If at least one (but not all) variable concerning 'Homework in <SUBJECT> lessons' is coded differently from 'Never' (1, 2, or 3), but the variable concerning 'Homework given in <SUBJECT> lessons' is coded to 'Never' (4), a warning has been given, but no corrections have been undertaken.

3111 (SG) This cleaning step identifies all inconsistency problems for <SUBJECT> 'The teacher gives us homework' and the dependent questions concerning homework in <SUBJECT> lessons. (Note:<MAT> refers to mathematics lessons <SCI> refers to science lessons).
→ If all variables concerning 'Homework in <SUBJECT> lessons' are coded differently from 'Never' (1, 2, or 3), but the variable concerning 'Homework given in <SUBJECT> lessons' is coded to 'Never' (4), then the variables concerning 'Homework given in <SUBJECT> lessons' has been coded to 'Invalid' (I).

3112 (SG) This cleaning step reports all inconsistencies between students' responses to 'Work in <SUBJECT> lessons in small groups' and 'Begin a new topic in <SUBJECT> by working in small groups'. Note: <MAT>

refers to mathematics lessons <SCI> refers to science lessons <BIO> refers to biology lessons <CHE> refers to chemistry lessons <EAR> refers to earth science lessons <PHY> refers to physics lessons
⊗ If group work never happens in <SUBJECT> (4), but group work is used when students begin a new topic in <SUBJECT>, a warning has been given, but no corrections have been undertaken.

3141 (SG) This cleaning step identifies problems with numbers of girls and boys within classes.
⊗ If the number of boys or girls is less than 10 percent of the number of all students in class under consideration, a warning has been given, but no changes have been undertaken.

INCONSISTENCY CHECK-TEACHER FILE

This set of cleaning steps identifies problems between data variables.

3201 (TG): This cleaning step identifies all inconsistencies in 'Number of boys in <CLASS>' and 'Number of girls in <CLASS>' (Note: <CLASS> can be the mathematics class or the science class. A/BTBMBOY and A/BTBMGIRL refer to the mathematics class. BTBSBOY and BTBSGIRL refer to the science class.)
→ If both variables were coded to 0, then both have been recoded to 'Invalid'.
⊗ If one variable was coded to 0 and the other was coded to 'Missing', a warning has been given, but no corrections have been undertaken.

3202 (TG): This cleaning step identifies inconsistencies in the variable block concerning different achievement levels in the<CLASS>. If the sum of all percentages exceeds 110, a warning is given. (Note: <CLASS> can be the mathematics class or the science class.)
⊗ If the sum of all percentages exceeds 110, a warning has been given, but no corrections have been undertaken.

3203 (TG): This cleaning step identifies inconsistencies in the lists concerning <SUBJECT> topics. Teachers were asked to indicate the number of periods a topic has been taught (<VAR>). They also should have indicated whether they will begin or continue teaching the topic (<VAR_A>), whether or not the topic is taught this year (<VAR_B>), and whether the topic was taught in a previous year (<VAR_C>). (Note: <SUBJECT> can be mathematics or science.)
→ If no periods were assigned to the topic (<VAR> 'Missing' (9)) and all other variables concerning the topic were coded to 'Not checked' (1), all 'Not checked' (1) have been recoded to 'Missing' (9).
→ If at least one variable was coded to 'Checked' (2), all variables coded to 'Missing' (9) have been recoded to 'Not checked' (1).
→ If the topic was taught a number of periods (<VAR> < 5), but it is indicated that the topic is not taught this year, (<VAR_B> 'Checked' (2)), <VAR_B> has been recoded to 'Not checked' (1).
⊗ If it is indicated that the topic is not taught this year (<VAR_B> 'Checked' (2)) and also that teaching the topic will be continued or begun (<VAR_A> 'Checked' (2)), then a warning has been given, but no corrections have been undertaken.

3231 (TG): This cleaning step identifies inconsistencies in the lists concerning <SUBJECT> topics. Teachers were asked to indicate the number of periods a topic has been taught (<VAR>). They also should have indicated whether they will begin or continue teaching the topic (<VAR_A>), whether or not the topic is taught this year (<VAR_B>), and whether the topic was taught in a previous year (<VAR_C>)
→ If the value for time assigned to the topic (<VAR>) was coded to 'Not administered' (8) and all other variables concerning the topic were coded to 'Not checked' (1), 'Checked' (2), or 'Missing' (9), the value for time assigned to the topic (<VAR>) has been recoded to 'Missing' (9).

3232 (TG): This cleaning step identifies inconsistencies in the lists concerning <SUBJECT> topics. Teachers were asked to indicate the number of periods a topic has been taught (<VAR>). They also should have indicated whether they will begin or continue teaching the topic (<VAR_A>), whether or not the topic is taught this year (<VAR_B>), and whether the topic was taught in a previous year (<VAR_C>). (Note: <SUBJECT> can be mathematics or science.)

⊗ If one (but not all) of the variables concerning the topic has been coded 'Not administered' (8), a warning has been given, but no corrections have been undertaken.

3204 (TG): This cleaning step identifies inconsistencies in the lists concerning <SUBJECT> topics. Teachers were asked to indicate the number of periods a topic has been taught. They also should have indicated whether they will begin or continue teaching the topic, if the topic is not taught, and if the topic was taught in a previous year. Note: <SUBJECT> can be mathematics or science.

→ If the respondent indicates that the topic was taught a number of periods, but all other variables concerning the topic were coded to 'Missing' (9), all 'Missings' have been recoded to 'Not checked' (1).

3205 (TG): This cleaning step identifies all inconsistencies between the sum of minutes assigned to teaching activities in <CLASS> and the duration of the <CLASS>.(Note: <CLASS> can be either the mathematics class or the science class).

⊗ If the total time assigned to activities exceeds the time available, a warning has been given, but no corrections have been undertaken.

3206 (TG): This cleaning step identifies all inconsistencies between order of teaching activities in <CLASS> and times assigned to teaching activities in <CLASS>. (Note: <CLASS> can be either the mathematics class or the science class).

→ If an activity was put in order, but 0 minutes were assigned to carrying out the activity, the time has been recoded to 'Invalid'.

→ If an activity was not put in order (0), but time was assigned to carrying out the activity, the order has been recoded to 'Invalid'.

3207 (TG): This cleaning step identifies inconsistencies in the variable blocks concerning the frequency with which different tasks are given as homework in <CLASS>.(Note: <CLASS> can be the mathematics class or the science class).

⊗ If the respondent indicates that more than zero tasks are assigned 'Rarely' to 'Always' and homework is never assigned for more than two tasks, a warning has been given, but no corrections have been undertaken.

→ If the respondent indicates for fewer than three tasks that they are assigned 'Never' (1) and for all other tasks that he/she does not assign homework and the variable indicating how often homework is assigned is coded to 'Never' (1), then 'Never' has been recoded to 'I do not assign homework' (5).

INCONSISTENCY CHECK-SCHOOL FILE

This set of cleaning steps identifies inconsistency problems between data variables.

3301 (CG): This cleaning step identifies all inconsistency problems with the grade levels found in a school.

→ If no grades can be found at a school and all grades are coded to 'No' (2), all variables have been recoded to 'Invalid' (I).

⊗ If 'Yes' (1) is not coded for any grade, but 'Missing' is found for at least one grade, a warning has been given, but no corrections have been undertaken.

3302 (CG): This cleaning step identifies all schools in which the grades in the school are not sequential.

⊗ A warning has been given, but no corrections have been undertaken.

3303 (CG): This cleaning step identifies all problems with the number of the classroom teachers indicated in Q. 3 & 4.

⊗ If the number of individual full-time classroom teachers is larger than the total number of full-time equivalent classroom teachers the observation is flagged, but no corrections have been undertaken

⊗ If the number of full-time classroom teachers and the number of part-time classroom teachers equals zero, a warning has been given, but no corrections have been undertaken.

3304 (CG): This cleaning step identifies all problems with the 'Percentage of the classroom teachers teach '. (Q.6).

⊗ If the sum of the percentage of teachers who teach more than three-quarters math (A/BCBMTEAC) and the percentage of teachers who teach no math (A/BCBMNONE) exceeds 100, a warning has been given, but no corrections have been undertaken.

⊗ If the sum of the percentage of teachers who teach more than three-quarters science (A/BCBSTEAC) and the percentage of teachers who teach no science (A/BCBSMNONE) exceeds 100, a warning has been given, but no corrections have been undertaken.

⊗ If the sum of the percentage of teachers who teach more than three-quarters math and science (A/BCBGTEAC) and the percentage of teachers who teach neither math nor science (A/BCBGNONE) exceeds 100, a warning has been given, but no corrections have been undertaken.

3305 (CG): This cleaning step identifies all problems with the total time the principal spends on activities (A/BCBGAC01-14).

⊗ If the sum of A/BCBGAC01-14 exceeds 280 hours per month, a warning has been given, but no corrections have been undertaken.

⊗ If the sum of A/BCBGAC01-14 equals zero, a warning has been given, but no corrections have been undertaken.

3306 (CG): This cleaning step identifies all problems with the question 'In your school, how many computers are ...'

⊗ If the total number of computers available for teachers and students (A/BCBGCOM1) is smaller than at least one of the numbers of computers used by teacher and students indicated in A/BCBGCOM2-4, a warning has been given, but no corrections have been undertaken.

3307 (CG): The listed observations have the sum of girls and boys for the whole school (A/BCBGGENR A/BCBGBENR), for the lower grade(A/BCBGLGER A/BCBGLBER) or for the upper grade (A/BCBGUGER A/BCBGUBER) equal to 0. The cases in which no children are enrolled in one target grade could be correct if the school is a lower grade or upper grade school only.

→ If the sum of boys and girls enrolled in a school equals zero, the number of boys and the number of girls enrolled have been recoded to 'Invalid' (9997).

→ If the sum of boys and girls enrolled in both target grades equals zero, the number of boys and the number of girls enrolled have been recoded to 'Invalid' (997) for both target grades.

⊗ If the sum of boys and girls enrolled in lower grades or in upper grade equals zero, but enrolled students are indicated for at least one target grade, the school has been flagged.

3308 (CG): This cleaning step identifies all problems with the number of girls/boys studying math or science compared to the number of girls/boys enrolled in the grade. It also identifies all problems with the number of girls/boys repeating a grade compared to the number of girls/boys enrolled in the grade.
⊗ If the number of students repeating a grade, studying math, or studying science is larger than the number of students enrolled in the grade, a warning has been given, but no corrections have been undertaken.

3309 (CG): This cleaning step identifies all cases in which the instructional time is the same for both the upper grade and the lower grades, and the times were incorrectly entered into the column for the lower grade instead of the column for the upper grade.
→ All values have been transcribed to the upper grade variables.

3310 (CG): This cleaning step identifies problems with the instructional times.
→ If the number of instructional days per school year is 0 for both target grades the corresponding variables A/BCBGLDYY and A/BCBGUDYY have been recoded to 'Invalid'.
→ If the number of total hours per week is 0 for both target grades the corresponding variables A/BCBGLTHW and A/BCBGUTHW have been recoded to 'Invalid'.
→ If the number of instructional hours per week is 0 for both target grades the corresponding variables A/BCBGLIHW and A/BCBGUIHW have been recoded to 'Invalid'. If A/BCBGINST indicates that the instructional time is the same for both grades, only the upper grade variables have been recoded.

3311 (CG) This cleaning step identifies all cases in which the sum of the percentage of students in the most advanced course and the least advanced course exceeds 110. The following variables have been checked:(1) most and least advanced course in math (A/BCBMUC41 A/BCBMUC42); (2) most and least advanced course in science (A/BCBMUC41 A/BCBMUC42).

⊗ A warning has been given, but no corrections have been undertaken.